Revolut : le compteur de tokens déplacé, pas supprimé

Revolut a fait d'un modèle chinois de Z.ai, exécuté sur ses propres serveurs, l'outil de code par défaut de quelque 2 000 ingénieurs, pour une facture déclarée dix fois inférieure. La presse y a lu une coupe de coûts et un revers pour Claude. L'épisode montre surtout que la forme de la facturation décide de ce qu'une organisation ose demander à l'IA, et que supprimer le compteur ne supprime pas le rationnement : il le déplace.

Par Alexandre Berge · 14 novembre 2026 · 13 min de lecture · Chronique

L'événement tel qu'il a été vu, à la date du 7 novembre 2026

Le 13 octobre 2026, sur la scène principale de la conférence AI Engineer à New York, David Abitbol, qui dirige les opérations d'IA et l'infrastructure de Revolut, projette deux courbes. Elles retracent, depuis janvier 2025, les délais d'un échantillon d'environ 2 000 ingénieurs de la banque numérique européenne. Le 3 novembre, The Register reprend l'intervention sous un titre qui fixe le récit : Revolut aurait divisé par dix le coût de son code assisté par IA en écartant Claude au profit de modèles qu'elle héberge elle-même.

La chronologie, telle que l'organisateur l'a publiée avec la transcription, compte trois temps. Gemini, de Google, arrive en janvier 2025. Claude Code, l'outil de programmation d'Anthropic, est déployé à grande échelle en janvier 2026 sous de strictes limites budgétaires, certains ingénieurs dépensant, selon David Abitbol, de 1 000 à 10 000 dollars par mois sans retour probant. En avril 2026, Revolut commence à auto-héberger des modèles à poids ouverts (les paramètres du modèle, publiés et téléchargeables, que l'entreprise exécute où elle veut), d'abord GLM 5.1 puis GLM 5.3, conçus par le laboratoire chinois Z.ai, anciennement Zhipu AI.

Tous les chiffres sont déclarés par l'intervenant, sans vérification indépendante publiée. Le passage à GLM 5.1 coïncide avec une amélioration de 60 % du temps de codage et de 70 % du délai de réalisation (de la prise en charge d'une fonctionnalité à son intégration dans le code prêt à déployer). Le service repose sur 56 GPU Nvidia B200 (des processeurs spécialisés dans le calcul des modèles). Le coût mixte avoisine 30 cents par million de tokens (les fragments de texte sur lesquels se mesure et se facture le travail d'un modèle), et la facture représente un dixième de ce que Revolut versait à Anthropic et à OpenAI. Claude Code et Codex restent disponibles, à titre d'exception.

Le récit dominant s'est construit sur trois registres : la coupe de coûts, l'abandon de Claude, la prouesse d'ingénierie qui consiste à servir un modèle d'environ 754 milliards de paramètres, selon The Register. Un quatrième affleure, la rivalité sino-américaine des modèles, que l'intervenant a lui-même abordée. Ce récit dit vrai sur les montants. Il ne dit presque rien de ce qui a changé dans le travail.

Le compteur fabriquait le rationnement qu'il devait seulement mesurer

Le passage le plus instructif de l'intervention porte sur un comportement. Tant que chaque token était compté, les ingénieurs le réservaient, selon David Abitbol, à la tâche la plus difficile. Sa formule : « You no longer have to save your tokens for the end of the quarter or the most difficult task » (vous n'avez plus à garder vos tokens pour la fin du trimestre ou pour la tâche la plus difficile).

L'économie comportementale a nommé ce mécanisme. Dans « Mental Accounting Matters » (1999), Richard Thaler montre que ménages et organisations regroupent leurs dépenses en catégories dotées de budgets, et qu'un budget épuisé freine une dépense même lorsque d'autres comptes disposent encore de fonds, parce que l'argent n'y est pas traité comme fongible. Un plafond de tokens par ingénieur et par période est exactement un tel compte. Il transforme une question de valeur (cette tâche vaut-elle d'être déléguée ?) en question de solde (reste-t-il assez de crédit ?). L'outil est gardé pour les cas jugés dignes de lui, et le travail ordinaire reste manuel.

Le compteur ne se contente donc pas de mesurer l'usage, il le sélectionne. Revolut avait de bonnes raisons de l'installer, la dépense étant élevée et le gain incertain. Mais en rendant chaque appel visible et coûteux pour celui qui le passe, la règle budgétaire a orienté l'outil vers une fraction étroite du travail. La synthèse de l'organisateur le précise : la comparaison interne n'isole pas l'effet du changement de modèle des autres causes possibles. Rien ne permet donc de conclure que Claude produisait moins de valeur par requête. Les chiffres suggèrent plutôt que la forme du contrat pesait au moins autant que la qualité du modèle.

Le token moins cher a élargi la demande au lieu de la réduire

Une fois le plafond levé, l'usage s'est déplacé. Le délai de réalisation s'améliore davantage que le temps de codage, et David Abitbol en conclut que le goulet d'étranglement de l'ingénierie n'est pas nécessairement le code. Le temps se perd dans la messagerie, dans Jira et Confluence (les outils de suivi de tickets et de documentation), auxquels The Register ajoute le débogage. Ces tâches périphériques, nombreuses et peu valorisées une à une, sont précisément celles qu'aucun budget de tokens n'aurait financées.

La demande, elle, a augmenté. Selon une diapositive de l'intervention, entre avril et août, à nombre de GPU constant, le volume de requêtes par seconde a progressé de 76 %, tandis que le délai de première réponse passait de 14,26 à 1,68 secondes.

> **Le paradoxe de Jevons (William Stanley Jevons, *The Coal Question*, 1865)**
>
> Au chapitre VII de *The Coal Question*, l'économiste britannique conteste l'idée que des machines à vapeur plus sobres ménageront les réserves de charbon du pays. « It is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption » (c'est une pure confusion d'idées que de supposer que l'usage économe du combustible équivaut à une consommation réduite). Une ressource dont le coût d'usage baisse trouve des emplois nouveaux, si bien que la consommation totale peut croître. L'économie de l'environnement a repris l'idée sous le nom d'effet rebond.

Transposé à Revolut, le *paradoxe de Jevons* éclaire ce que le récit de la coupe de coûts masque. Le prix unitaire du token a baissé d'un facteur dix, selon l'intervenant, et la ressource s'est répandue vers des emplois que son ancien prix excluait. **L'économie ne se lit pas seulement dans la facture, elle se lit dans l'extension du périmètre.** Fidèle à Jevons, le paradoxe porte sur la quantité de ressource consommée (calcul et tokens), non sur la dépense : que la facture déclarée ait baissé ne l'invalide pas. Un facteur absent du charbon victorien borne toutefois le rebond. La flotte de GPU est fixe.

Le rationnement change de forme, de la facture à la file d'attente

C'est le point que le récit dominant laisse dans l'ombre. Avec 56 GPU, la capacité n'est pas illimitée, elle est seulement payée d'avance. David Abitbol en a donné les bornes. Au-delà de cinq secondes avant le premier token, sa messagerie Slack se remplit de plaintes, d'où une cible inférieure à trois secondes. Une vitesse de 200 tokens par seconde suffit, et les ingénieurs qui se connectent après 17 heures, quand les GPU sont moins sollicités, en obtiennent 800. La fenêtre de contexte (la quantité de texte que le modèle prend en compte à chaque requête) est volontairement bornée à 256 000 tokens alors que le modèle peut en servir un million. « Long context is a scam » (le contexte long est une arnaque), a lancé l'intervenant, qui conseille de résister à l'ajout de GPU tant que la configuration n'a pas été réglée.

Chacun de ces paramètres est une règle d'allocation. Le prix ne rationne plus, la latence, l'horaire et la taille du contexte le font à sa place. **Le compteur n'a pas disparu, il est passé de la direction financière à l'équipe d'infrastructure**, et son unité n'est plus le dollar mais la seconde d'attente. Une file d'attente ne culpabilise pas, elle ralentit, mais elle reste une rareté administrée.

AOL, 1996 : l'illimité rencontre le nombre fini de modems

L'histoire des réseaux offre un cas d'école de ce déplacement. En décembre 1996, America Online, fort de 7 millions d'abonnés, remplace sa facturation horaire par un forfait illimité à 19,95 dollars par mois. Le Deseret News relevait que la société s'attendait à voir s'envoler le

temps passé en ligne et ajoutait des équipements. L'afflux dépasse pourtant la capacité, et des milliers d'abonnés se heurtent à la tonalité occupée de leur modem.

Le 29 janvier 1997, AOL conclut avec l'État de Washington et 35 autres États un accord volontaire : remboursements modulés selon le temps de connexion effectif, résiliation facilitée, publicité suspendue en février tant que les problèmes de réseau nés du forfait ne sont pas résolus. La société nie expressément toute faute. La procureure générale Christine Gregoire résume : « AOL's marketing clearly got out ahead of its technological capacity » (le marketing d'AOL a nettement devancé sa capacité technologique).

Le parallèle tient par la structure. La suppression du coût marginal (le prix de l'unité supplémentaire consommée) libère une demande comprimée par le compteur, et le rationnement se reporte sur la capacité. AOL rationnait par la tonalité occupée, Revolut par le délai de réponse. La différence tient à la préparation : Revolut a dimensionné et réglé sa flotte pour une population interne de taille connue.

Internaliser l'inférence échange une dépendance contre une autre

Le second angle mort concerne la dépendance. En internalisant l'inférence (l'exécution du modèle pour produire une réponse), Revolut cesse de dépendre des prix et des plafonds d'Anthropic et d'OpenAI pour son usage courant. Elle dépend désormais des publications et des licences de Z.ai, de la disponibilité des GPU et de compétences internes de réglage, sans lesquelles, selon l'intervenant, les performances sont médiocres.

Les licences l'illustrent. GLM 5.1 est publié sous licence MIT, l'une des plus permissives. GLM 5.3 l'est sous une licence propre à Z.ai, ouverte pour l'essentiel, qui ajoute une condition : une entreprise exploitant une activité de modèle en tant que service (la fourniture à des tiers d'un accès à l'inférence) et réalisant plus de 10 milliards de dollars de chiffre d'affaires sur douze mois doit passer un examen de sécurité dont Z.ai fixe la portée. L'usage interne de Revolut n'entre pas dans cette définition telle qu'elle est rédigée. Le passage d'une version à l'autre montre toutefois que les conditions d'accès aux poids évoluent au gré de l'éditeur. Par ailleurs, Beijing Zhipu Huazhang Technology, connue sous le nom de Zhipu AI, figure depuis le 16 janvier 2025 sur l'Entity List du Bureau of Industry and Security américain. L'inscription encadre les exportations vers l'entreprise, non l'usage de poids déjà publiés, mais elle signale un fournisseur exposé aux aléas d'une politique commerciale.

David Abitbol répond par le chemin des données : aucun point d'accès situé en Chine, des agents de code encadrés par un bac à sable (un environnement d'exécution cloisonné), des règles d'approbation humaine et une télémétrie détaillée. La synthèse de l'organisateur, qui y voit sa lecture de la situation de Revolut et non une règle générale, conclut que l'auto-hébergement laisse Revolut responsable de l'usage, du contrôle et de l'enquête en cas d'incident. **La charge ne disparaît pas avec la facture, elle migre vers le run** (l'exploitation courante), l'infrastructure et la conformité.

Le mécanisme sans la marque : quand la forme du budget dessine l'usage

Retiré de son contexte, le mécanisme tient en trois enchaînements. Un outil facturé à l'unité, de valeur incertaine, appelle un plafond. Le plafond crée un compte mental qui réserve l'outil aux tâches jugées nobles et laisse en friche le travail périphérique, souvent le plus chronophage. Lorsque le coût marginal tombe, par internalisation ou par forfait, l'usage s'étend à ce travail, la demande croît, et le rationnement réapparaît sous forme de contrainte de capacité, gérée par d'autres acteurs que ceux qui tenaient le budget.

Trois conditions suffisent à le reproduire : une dépense à l'usage dispersée entre de nombreux utilisateurs, un gisement de tâches à faible valeur unitaire mais à fort volume, une organisation capable d'acquérir une capacité fixe. La chronique « Facture IA de l'État : la tentation de rationner l'agent », publiée le 17 octobre 2026, décrivait la tentation du plafonnement face à un outil gratuit au point d'usage. Le cas Revolut en montre l'envers et le prix.

Avant l'internalisation, le signe le plus net est une consommation concentrée à l'approche de l'échéance budgétaire. Après, ce sont les plaintes de latence, l'usage reporté aux heures creuses et les demandes d'extension de flotte.

> **Grille de lecture**
>
> 1. Les outils d'IA de l'organisation sont-ils soumis à des plafonds par personne ou par période, et la consommation se concentre-t-elle à l'approche de leur échéance ?
> 2. Quelle part des requêtes porte sur le travail périphérique (tickets, documentation, messagerie, débogage) plutôt que sur la tâche principale ?
> 3. Quel indicateur la direction suit-elle pour ces outils, la dépense ou le délai de réalisation du travail ?
> 4. En cas de capacité interne, qui fixe les règles de latence, de priorité et de taille de contexte, et ces règles sont-elles connues des utilisateurs ?
> 5. Quels engagements de licence lient l'organisation au fournisseur des poids, et à quelle date ont-ils été relus ?

La capacité réglera ce que le prix réglait

Revolut n'a pas supprimé le compteur, elle l'a remplacé. Le token presque gratuit pour l'ingénieur a ouvert des usages que le budget excluait, comme Jevons le décrivait pour le charbon, et la rareté s'est reportée sur 56 GPU administrés par une équipe d'infrastructure, comme elle s'était reportée sur les modems d'AOL. La coupe de coûts est la cause, le déplacement des usages est l'effet durable.

Une hypothèse en découle, vérifiable de l'extérieur d'ici au 30 juin 2028. Si l'effet Jevons opère, Revolut devrait soit annoncer une extension de sa flotte au-delà des 56 B200, soit

décrire publiquement des règles de priorité ou de quota sur sa capacité interne, et au moins une autre grande organisation régulée européenne devrait annoncer une internalisation comparable en invoquant la facturation forfaitaire. Si la demande plafonne à capacité constante, le rebond aura été borné, et le cas relèvera d'un arbitrage de prix plus que d'une transformation des usages.

Note méthodologique

Les faits relatifs à Revolut reposent sur la page de l'intervention publiée par l'organisateur de la conférence, avec sa transcription, et sur l'article de The Register du 3 novembre 2026. Tous les chiffres (coûts, gains de délai, demande) sont déclarés par l'intervenant, non audités, et donc attribués. Le lien entre levée des plafonds et gains est celui qu'avance David Abitbol, et la synthèse de l'organisateur rappelle qu'il n'est pas isolé des autres causes. Divergence relevée : The Register associe le coût de 30 cents à GLM 5.1 et place le débogage parmi les gains, la transcription parle du coût mixte de l'auto-hébergement de GLM et cite la messagerie, Jira et Confluence. Le choix du *paradoxe de Jevons* comme cadre central, associé à la *comptabilité mentale* de Thaler, et celui d'AOL comme précédent relèvent de l'auteur. Aucune communication propre de Revolut, d'Anthropic ou de Z.ai sur cet épisode n'a été relevée dans la presse et les sites consultés. Faits arrêtés au 7 novembre 2026.

Recherche de sources, croisement des données et structuration du raisonnement s'appuient sur une assistance par intelligence artificielle. Le choix de l'angle et du cadre théorique, la vérification des sources, les arbitrages et l'établissement du texte final relèvent de l'auteur.

Sources

Documents des organisations et décisions

AI Engineer, « Shipping like a startup, regulated like a bank », intervention de David Abitbol (Revolut), AI Engineer NYC 2026, 13 octobre 2026, page et transcription consultées le 7 novembre 2026, https://ai.engineer/nyc/2026/talks/revolut

Z.ai (zai-org), « GLM-5.1, LICENSE » (MIT License), Hugging Face, consulté le 7 novembre 2026, https://huggingface.co/zai-org/GLM-5.1/blob/main/LICENSE

Z.ai (zai-org), « GLM-5.3, LICENSE » (GLM-5.3 License), Hugging Face, consulté le 7 novembre 2026, https://huggingface.co/zai-org/GLM-5.3/blob/main/LICENSE

Bureau of Industry and Security, Department of Commerce, « Addition of Entities to and Revision of Entry on the Entity List », Federal Register, vol. 90, n° 10, 90 FR 4617, 16 janvier 2025, https://www.govinfo.gov/content/pkg/FR-2025-01-16/pdf/2025-00704.pdf

Presse de référence

Tobias Mann, « British fintech Revolut slashes AI coding costs 10x by cutting out Claude and going self-hosted », The Register, 3 novembre 2026, https://www.theregister.com/ai-and-ml/2026/11/03/british-fintech-revolut-slashes-ai-coding-costs-10x-by-cutting-out-claude-and-going-self-hosted/5307345

Cadres théoriques

William Stanley Jevons, The Coal Question, chapitre VII « Of the Economy of Fuel », 1865, édition Library of Economics and Liberty consultée le 7 novembre 2026, https://www.econlib.org/library/YPDBooks/Jevons/jvnCQ.html?chapter_num=9

Richard H. Thaler, « Mental Accounting Matters », Journal of Behavioral Decision Making, vol. 12, n° 3, p. 183-206, septembre 1999, https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-0771%28199909%2912%3A3%3C183%3A%3AAID-BDM318%3E3.0.CO%3B2-F

Parallèle historique

Brooke Adams, « AOL's flat rate is a good deal offered in an insensitive way », Deseret News, 2 décembre 1996, https://www.deseret.com/1996/12/2/19280328/aol-s-flat-rate-is-a-good-deal-offered-in-an-insensitive-way/

Washington State Office of the Attorney General, « Washington reaches agreement with America Online », 29 janvier 1997, https://www.atg.wa.gov/news/news-releases/washington-reaches-agreement-america-online-0

Chronique liée

Alexandre Berge, « Facture IA de l'État : la tentation de rationner l'agent », elegartech.fr, 17 octobre 2026, https://elegartech.fr/articles/facture-ia-de-l-etat-la-tentation-de-rationner-l-agent