

AU SOMMAIRE

5 actus · 1 Décryptage

- 01 Uber supprime 3 300 postes pour financer les voitures sans chauffeur
- 02 Faire travailler une IA en continu coûte jusqu'à 45 % moins cher
- 03 New York interdit l'IA à plus d'un demi-million d'élèves
- 04 Meta éteint pour de bon la caméra de milliers de lunettes trafiquées
- 05 Une IA dessine un logiciel image par image, sans une ligne de code

LE DÉCRYPTAGE

Grand modèle de langage (Large Language Model) : ce qui se passe vraiment quand une IA vous répond.



01

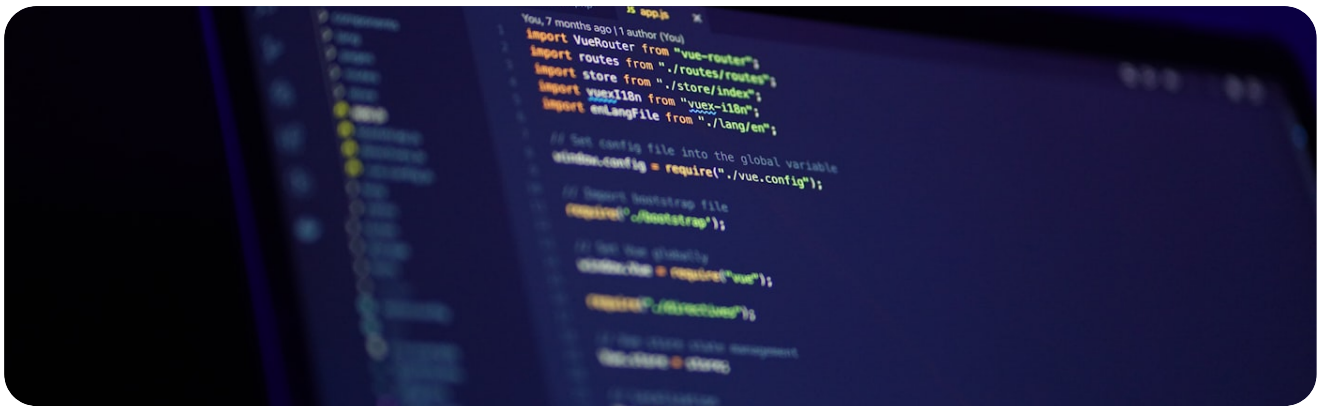
Uber supprime 3 300 postes pour financer les voitures sans chauffeur

Le service de réservation de trajets se sépare de 10 % de ses salariés dans le monde. L'argent économisé doit partir vers les taxis autonomes.

Uber a annoncé mercredi la suppression d'environ 3 300 postes, soit 10 % de ses effectifs mondiaux. C'est sa plus grosse coupe depuis 2020, et elle ne répond pas à une chute de la demande de trajets. L'entreprise dit vouloir moins d'étages hiérarchiques, des décisions plus rapides et moins de doublons entre équipes. Le nombre de managers baisse d'environ 20 %, et l'effectif revient autour de 30 000 personnes, son niveau de 2021.

La destination de l'argent est assumée. Uber a déjà engagé plus de 10 milliards de dollars dans des partenariats de robotaxi (des voitures qui circulent sans conducteur). L'entreprise se prépare à un marché où le chauffeur n'est plus au centre du modèle, et préfère payer cette bascule en réduisant ses frais de structure plutôt qu'en allant chercher de l'argent à l'extérieur.

Le signal utile n'est pas technologique, il est comptable. Uber ne supprime pas des postes parce que des machines font déjà le travail : il supprime des postes pour avoir de quoi payer le moment où elles le feront. C'est un arbitrage de trésorerie, pas un constat de productivité. La question qu'il pose à toute entreprise est simple : quelle part de vos frais fixes finance encore l'organisation d'hier plutôt que le produit de demain ?



02

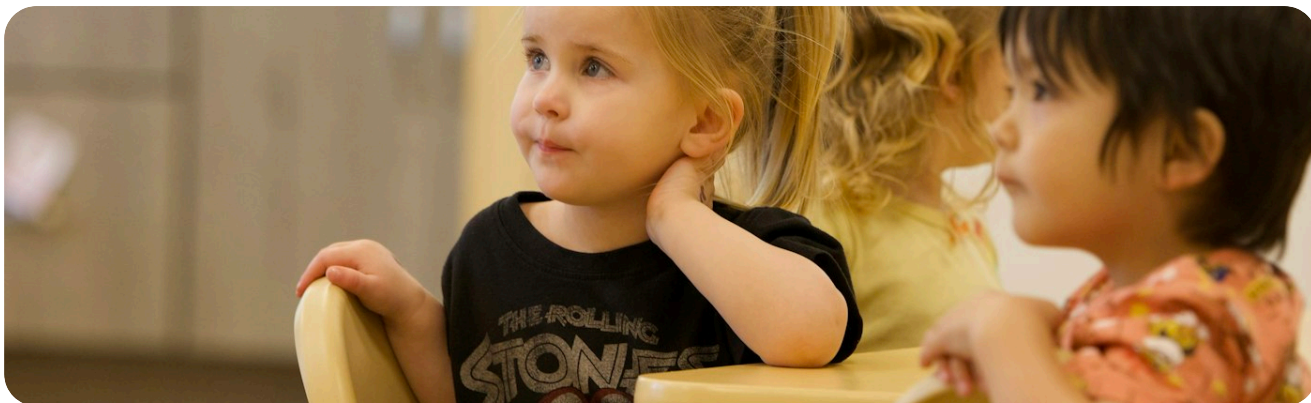
Faire travailler une IA en continu coûte jusqu'à 45 % moins cher

Anthropic a publié deux nouvelles versions de Claude. Le prix des tâches longues, celles où l'IA enchaîne des dizaines d'étapes seule, baisse fortement.

Anthropic a mis en ligne Claude Fable 5.1 et Claude Mythos 5.1. C'est le même modèle avec deux niveaux de garde-fous : Fable 5.1 est ouvert à tous, Mythos 5.1 reste réservé à des programmes d'accès contrôlé pour la cybersécurité et les sciences du vivant. Le tarif affiché est de 10 dollars par million de jetons (les fragments de texte dans lesquels l'IA compte tout) envoyés, et 50 dollars par million produits.

La vraie nouveauté est ailleurs. La relecture du cache, cette mémoire tampon où l'IA garde ce qu'elle a déjà lu pour ne pas le repayer à chaque étape, tombe à 0,25 dollar par million de jetons, soit 75 % de moins qu'avant. Anthropic annonce environ 25 % d'économie sur un usage courant, et jusqu'à 45 % sur les tâches où l'IA travaille seule pendant des heures.

C'est précisément ce poste de dépense qui décide si une automatisation passe du test au quotidien. Une chaîne qui coûtait 400 euros par mois à faire tourner en continu tombe autour de 220 euros, sans qu'une ligne de votre paramétrage change. Ce qui bouge, c'est le seuil à partir duquel l'idée devient rentable. Il vaut la peine de ressortir les projets abandonnés il y a six mois pour facture trop lourde : certains sont redevenus payants sans que vous fassiez rien.



03

New York interdit l'IA à plus d'un demi-million d'élèves

Le plus grand réseau d'écoles publiques des États-Unis coupe l'accès aux assistants IA jusqu'à la fin du collège. L'usage ne reprend qu'au lycée, et sous conditions.

Les écoles publiques de New York ont ouvert l'année scolaire 2026-2027 avec la restriction la plus large du pays : plus aucun accès à l'IA générative, ces outils qui rédigent du texte à la demande, pour les élèves de la maternelle jusqu'à la dernière année de collège, soit environ 14 ans. Plus d'un demi-million d'élèves sont concernés. Les logiciels qui embarquent un assistant destiné aux élèves sont retirés, tuteurs automatiques et robots de conversation compris.

Trois arguments reviennent chez les enseignants et les parents : la santé mentale, le développement des capacités d'attention et de raisonnement, et le coût environnemental. L'IA ne réapparaît qu'au lycée, de façon encadrée, surtout pour apprendre comment la technologie fonctionne. Trois ans plus tôt, la même académie avait bloqué ChatGPT, puis levé l'interdiction en expliquant qu'elle avait sous-estimé le potentiel de l'outil.

Ce mouvement de balancier vaut d'être observé au-delà de l'école. Une grande organisation ouvre, mesure, puis referme partiellement : la même séquence commence à se jouer dans la santé, le droit et la banque. Pour qui vend un produit à ces secteurs, un critère d'achat apparaît : savoir désactiver proprement l'IA de son outil, par profil d'utilisateur, sans casser le reste. C'est souvent ce détail qui décide du renouvellement du contrat.



04

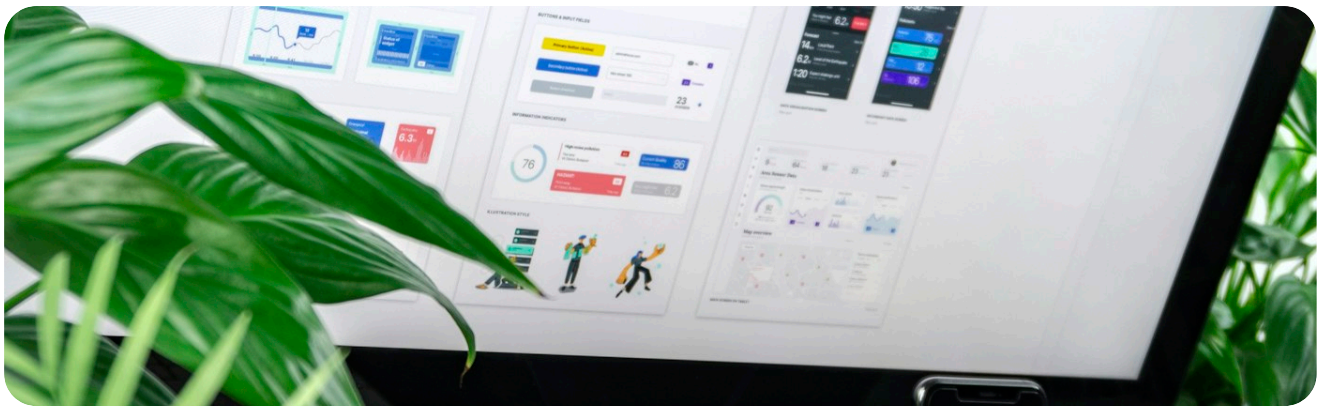
Meta éteint pour de bon la caméra de milliers de lunettes trafiquées

Des propriétaires avaient fait retirer la petite lumière qui signale l'enregistrement. Meta a coupé la caméra à distance, sans retour possible.

Les lunettes connectées de Meta, vendues sous les marques Ray-Ban et Oakley, portent une petite diode blanche qui s'allume dès que la caméra enregistre. C'est le seul signal dont dispose la personne en face. Depuis 2024, masquer cette lumière avec du ruban adhésif suffit à couper la caméra. Des bricoleurs contournaient le contrôle en retirant physiquement la diode, un service facturé environ 60 dollars.

Meta a répondu par une mise à jour obligatoire : si le système détecte que la diode est masquée, abîmée ou absente, la caméra s'arrête. L'entreprise indique avoir désactivé de cette façon la caméra de plusieurs milliers de paires. Un autocollant retiré rend la fonction ; une diode dessoudée, non : la caméra est perdue définitivement. Meta dit aussi faire supprimer les annonces qui vendent ces modifications.

Le précédent dépasse la question de la vie privée. Un fabricant vient de démontrer qu'il peut retirer à distance une fonction d'un appareil déjà payé, parce que l'usage sortait du cadre prévu. Pour qui achète du matériel connecté, caméras, capteurs, bornes ou lunettes, la question à poser au fournisseur n'est plus seulement le prix ni la garantie. C'est ce qu'il conserve le droit d'éteindre après la vente, et dans quels cas.



05

Une IA dessine un logiciel image par image, sans une ligne de code

Runway a présenté un système qui fabrique l'écran d'une application en direct, pendant qu'on l'utilise. Derrière, il n'y a rien : pas de programme, seulement des images calculées.

Runway, connu pour ses outils de vidéo générée, a dévoilé Solaris. Le principe rompt avec la façon habituelle de fabriquer un logiciel : au lieu d'écrire un programme qui affiche des boutons, l'IA produit directement l'image de l'écran, puis l'image suivante à chaque clic ou glissement de souris. Comptez moins d'une demi-seconde par image, en définition 720p.

Dans un test auprès de 250 personnes, les participants ont préféré Solaris à une interface écrite en code dans 61 % des cas pour le respect des consignes données, et 71 % des cas pour le naturel du comportement. Runway assume les limites : le texte affiché reste imparfait, l'accessibilité n'est pas réglée, et l'écran perd le fil sur les longues sessions. C'est présenté comme un aperçu de recherche, pas comme un produit à vendre.

L'intérêt immédiat se situe avant le produit fini. Une maquette cliquable classique ne réagit que là où quelqu'un a prévu qu'elle réagisse ; ici l'écran improvise, ce qui permet de mettre un vrai utilisateur devant une idée en quelques heures et de voir où il se perd. Si la méthode tient sur la durée, fabriquer une application changera de nature : on décrira le comportement attendu, plus la structure technique qui le produit.

Grand modèle de langage (Large Language Model)

Aussi dit LLM · modèle de langage

Un grand modèle de langage est un programme qui a lu une quantité colossale de textes et qui, à partir de ce que vous lui écrivez, calcule le mot le plus probable pour continuer, puis le suivant, jusqu'à former une réponse entière.

EN UNE IMAGE

Le clavier de votre téléphone qui vous propose le mot suivant quand vous tapez un message. Un grand modèle de langage, c'est le même mécanisme, entraîné sur une bibliothèque immense et capable de tenir la distance sur plusieurs pages au lieu de trois mots.

Votre phrase → mot suivant le plus probable → et encore → une réponse entière

ANALOGIE (POUR UN ENFANT DE 12 ANS)

Imaginez un camarade de classe qui aurait lu tous les livres de la bibliothèque de l'école. Vous commencez une phrase, il la termine, et le plus souvent il tombe juste : il a vu des milliers de phrases qui démarraient pareil. Mais il ne va jamais rouvrir un livre pour vérifier, et il ne dit presque jamais « je ne sais pas ». Alors quand il ne sait pas, il termine quand même la phrase, avec exactement le même aplomb.

Pourquoi ça compte : presque tout ce qui surprend dans les outils d'IA découle de cette mécanique. Le modèle complète, il ne consulte pas. Trois conséquences très concrètes : il écrit très bien même quand il a tort, parce que la qualité de la phrase ne dépend pas de sa véracité ; il ignore tout ce qui s'est passé après son entraînement, donc votre tarif d'aujourd'hui ; et il se facture au volume de texte, puisqu'il travaille morceau par morceau. La décision que ça débloque tient en une phrase : on ne lui demande pas une information vérifiable sans lui fournir le document qui la contient. On lui demande de la reformulation, du tri, du résumé, de la mise en forme.

Exemple en entreprise : une boutique en ligne de quatre personnes fait rédiger 200 fiches produits. En une matinée, le texte sort propre, dans le bon ton, avec la même structure partout. Mais sur une dizaine de fiches, le modèle a ajouté une garantie de cinq ans que la boutique n'a jamais proposée : la phrase était plausible, il l'a donc écrite. La règle adoptée ensuite est simple : les caractéristiques chiffrées, dimensions, garantie, prix, sont collées depuis le fichier fournisseur, et le modèle ne s'occupe que de la rédaction autour. Résultat : deux jours de travail économisés, et plus aucune promesse inventée.

À retenir : un grand modèle de langage ne sait pas, il complète. Il produit la suite la plus plausible, pas la plus vraie.

À NE PAS CONFONDRE

ChatGPT, Claude, Gemini : ce sont des produits, pas des modèles. Le modèle est le moteur ; l'application autour ajoute la mémoire, l'accès au web, les outils et les garde-fous. Changer de modèle à l'intérieur du même produit change la facture et la qualité des réponses, sans rien changer à votre façon de travailler.

Un moteur de recherche : il vous renvoie des pages qui existent déjà et que vous pouvez ouvrir. Un modèle de langage fabrique une phrase nouvelle, qui peut ne correspondre à aucune page. C'est pour cette raison qu'une réponse d'IA sans lien n'est pas une source.