KLEAA

Lire en ligne



OpenAI rattrape Claude en baissant ses prix

Bon jeudi à tous.

OpenAI baisse de 20 % le prix de sa meilleure IA et revient au contact de Claude. À Washington, l’administration Trump envisage d’obliger les entreprises à dire publiquement si elles utilisent une IA chinoise. En France, OVHcloud publie aujourd’hui sa première IA, assez légère pour tourner sur un simple ordinateur. HeyGen fabrique des vidéos de près de 30 minutes à partir d’une seule photo, et Apple lance une étude sur 200 000 personnes pour comprendre quand on fait confiance à l’IA sur sa santé. En fin d’édition, nous décryptons les modèles de raisonnement, ces IA qui réfléchissent avant de répondre.

Quelqu’un devrait lire ça ?

AU SOMMAIRE

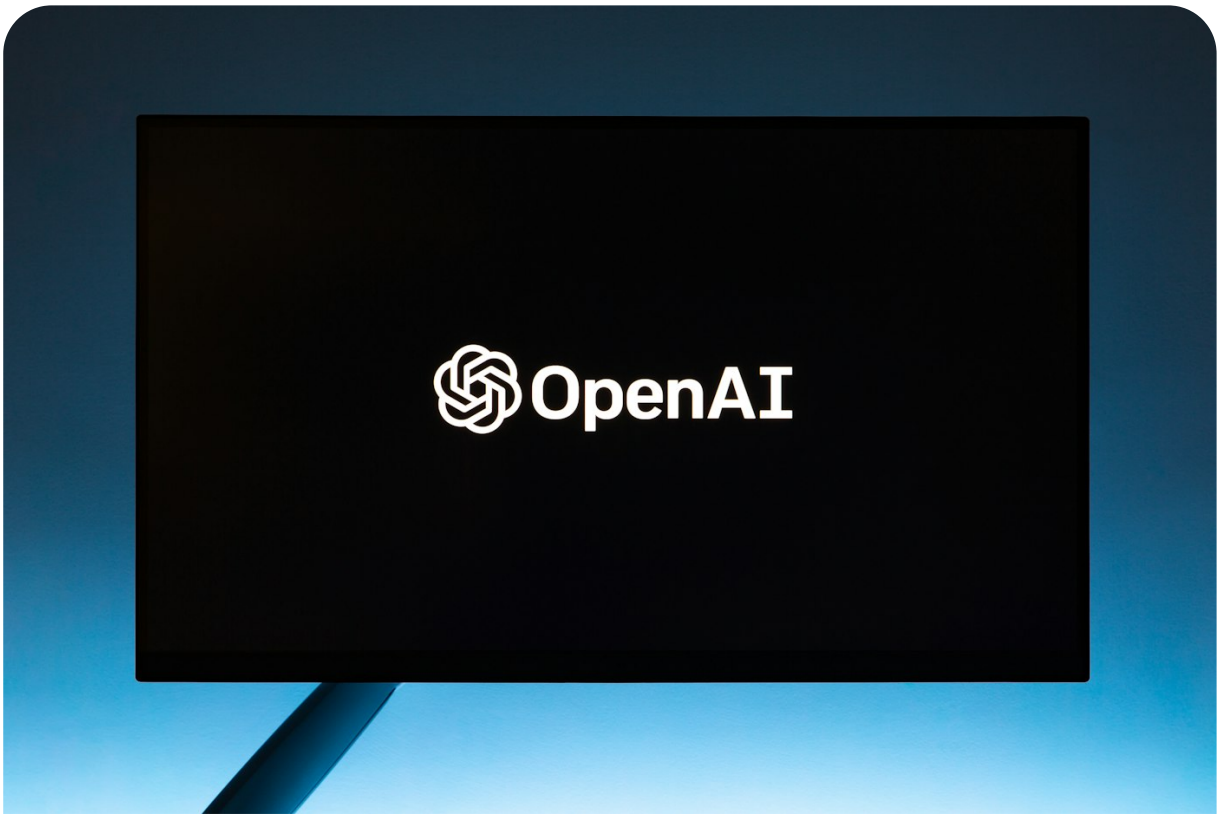
01 OpenAI baisse de 20 % le prix de sa meilleure IA et talonne Claude

02 Washington veut obliger les entreprises américaines à dire si elles utilisent une IA chinoise

03 OVHcloud publie aujourd’hui sa première IA, assez légère pour tourner sur un simple ordinateur

04 HeyGen fabrique des vidéos de près de 30 minutes à partir d’une seule photo

05 Apple lance une étude sur 200 000 personnes pour savoir si l’on peut faire confiance à l’IA sur sa santé



01

OpenAI baisse de 20 % le prix de sa meilleure IA et talonne Claude

Sorties mardi, GPT-6.2 Astra et GPT-6.2 Sol arrivent dans ChatGPT pour les abonnés payants. Astra se classe juste derrière Claude Fable 5.2 dans les évaluations indépendantes, pour beaucoup moins cher.

OpenAI a lancé mardi 2 nouvelles versions de son IA. GPT-6.2 Astra vise le travail exigeant (sciences, droit, finance), GPT-6.2 Sol les tâches rapides du quotidien. Elles sont disponibles dans ChatGPT pour les abonnés Plus, Pro, Entreprise et Éducation, et pour les développeurs. Pour ces derniers, Astra coûte 8 dollars par million de jetons envoyés et 40 dollars par million de jetons reçus (un jeton est un morceau de mot, l'unité dans laquelle l'IA est facturée), soit 20 % de moins que la version précédente. Sol garde son prix : 2 et 10 dollars.

Sur la qualité, les mesures indépendantes sont plus nuancées que les annonces. Le classement d'Artificial Analysis, un cabinet qui teste les IA, place Astra 2e avec 62 points, derrière Claude Fable 5.2 d'Anthropic (65) et devant Gemini 4 Carbon de Google (60). Un autre institut, Epoch AI, met Astra et Fable à égalité. Mais en droit, Astra n'arrive que 11e d'un test spécialisé, et sur une série de questions volontairement très difficiles, il invente une réponse dans 59 % des cas au lieu d'admettre qu'il ne sait pas.

OpenAI reconnaît aussi que ces modèles atteignent son niveau de risque « critique » en cybersécurité : ils pourraient aider sérieusement à pirater un système. L'entreprise a donc durci ses garde-fous, au risque de bloquer aussi des demandes légitimes. Pour vous, l'essentiel est ailleurs : la course entre OpenAI et Anthropic se joue désormais autant sur le prix que sur l'intelligence, et ce sont les utilisateurs qui en profitent.



02

Washington veut obliger les entreprises américaines à dire si elles utilisent une IA chinoise

Selon l’agence Reuters, l’administration Trump étudie une règle qui imposerait d’afficher sur son site web l’usage de modèles d’IA chinois. Sans les interdire, mais pour les décourager.

Le Trésor américain et le bureau scientifique de la Maison-Blanche en ont discuté. L’idée : une entreprise qui utilise une IA conçue en Chine, comme celles d’Alibaba ou de DeepSeek, devrait l’indiquer publiquement sur son site. Rien ne serait interdit, mais chacun pourrait voir qui s’en sert. La liste exacte des modèles concernés n’est pas connue.

La pression a déjà commencé. 2 élus républicains ont écrit à Airbnb et à DoorDash (livraison de repas) pour leur demander des comptes sur leur usage de modèles chinois. Airbnb répond qu’il s’appuie surtout sur des IA américaines et n’utilise que quelques modèles chinois gratuits, hébergés chez des prestataires américains agréés. D’autres pistes sont sur la table : limiter les achats publics, ou placer certains laboratoires chinois sur une liste noire du commerce américain.

Pourquoi tant d’entreprises s’en servent-elles ? Parce que beaucoup de modèles chinois sont à poids ouverts (téléchargeables et utilisables gratuitement), performants et peu chers. Rien n’est décidé et le texte peut encore changer. Mais le message est clair : le choix d’une IA devient aussi un choix politique, que vos clients et vos partenaires pourraient vous demander de justifier.

Le regard KLEAA

Cette règle ne vous concerne pas directement si vous travaillez en Europe, mais elle annonce une question qui arrivera vite chez vous : quelle IA tourne derrière vos outils ? Beaucoup de logiciels intègrent un modèle sans le dire. Prenez l’habitude de poser la question à vos fournisseurs, et notez la réponse. Le jour où un grand client ou un appel d’offres vous la posera, vous serez prêt.

Un modèle chinois installé sur un serveur européen, qui n’envoie rien en Chine, est-il encore une IA chinoise ?



03

OVHcloud publie aujourd’hui sa première IA, assez légère pour tourner sur un simple ordinateur

L’hébergeur français met en ligne Olala mini, gratuit et utilisable par les entreprises. Le moteur de recherche Qwant s’en sert déjà.

OVHcloud, l’un des grands hébergeurs européens de sites et de données, installé à Roubaix, a dévoilé mardi son premier modèle de langage (une IA qui comprend et écrit du texte). Baptisé Olala mini, il est publié ce jeudi en téléchargement libre, sous une licence qui autorise l’usage commercial et la modification. Après Mistral, OVHcloud devient la 2e entreprise française à construire ses propres modèles.

Sa particularité, c’est sa petite taille. Il compte 7 milliards de paramètres (les réglages internes qui font son intelligence), dont seulement 1 milliard travaille à chaque étape. Résultat : selon OVHcloud, il tourne avec une carte graphique de 8 Go, sur un ordinateur ou un petit serveur, sans envoyer vos données à l’extérieur. Il est pensé pour les tâches simples : résumer, répondre à des questions sur des documents internes, trier des messages. Qwant l’utilise pour ses réponses éclair et annonce 20 % de vitesse en plus, pour un coût divisé par 3.

Prudence toutefois : aucun score de qualité n’a encore été publié, et le modèle peine sur les raisonnements complexes et les longs documents. Son entraînement a coûté environ 7 millions d’euros de calcul, sur 3 supercalculateurs publics européens. Une version bien plus grosse, au-delà de 100 milliards de paramètres, est promise. Pour une PME qui veut une IA chez elle, sur des données sensibles, c’est une option européenne de plus à tester.



04

HeyGen fabrique des vidéos de près de 30 minutes à partir d’une seule photo

Avec Avatar 6, présenté mardi, le spécialiste des avatars vidéo crée des présentateurs numériques capables de parler d’une seule traite pendant une demi-heure.

HeyGen fabrique des avatars, c’est-à-dire des doubles numériques qui parlent face caméra à partir d’un texte ou d’un enregistrement audio. Sa nouvelle version, Avatar 6, n’a besoin que d’une photo pour créer un personnage, avec en option un court enregistrement pour affiner la ressemblance. Les vidéos de démonstration durent de 25 à 28 minutes, d’un seul tenant, sans boucle visible, dans une image plus fine que la haute définition classique.

Les nouveautés visent les usages professionnels. On peut demander un geste précis sur un mot précis du texte, placer 2 personnes dans la même image pour un faux podcast, ou corriger un mot dans une vidéo déjà produite : seul ce passage est refait, avec la même voix, le reste ne bouge pas. HeyGen n’a pas communiqué le prix de cette version.

Pour limiter les abus, un avatar créé à partir des vidéos d’une vraie personne exige une déclaration de consentement enregistrée par celle-ci, avec un code à usage unique, vérifiée avant toute création. Pour une petite entreprise, cela ouvre la porte à des formations, des tutoriels produits ou des messages clients réalisés sans studio ni tournage.

Le regard KLEAA

La vidéo d’entreprise change d’économie : ce qui demandait un tournage, un monteur et une demi-journée se fait désormais depuis un bureau. C’est une vraie occasion de former vos équipes ou d’expliquer vos produits. Mais un avatar ne crée pas la même confiance qu’un visage réel. Réservez-le aux contenus d’information, et gardez la vraie caméra pour ce qui engage la relation, comme un message du dirigeant ou une réponse à un client mécontent. Et dites clairement quand une vidéo est générée : vos clients finiront par le voir.



05

Apple lance une étude sur 200 000 personnes pour savoir si l’on peut faire confiance à l’IA sur sa santé

Avec OpenEvidence, une IA médicale utilisée par les soignants, Apple va observer pendant plus d’un an comment des Américains réagissent aux réponses d’une IA sur leur santé.

L’étude a été annoncée le 15 novembre. Jusqu’à 200 000 adultes américains équipés d’un iPhone récent pourront poser leurs questions de santé du quotidien dans une application dédiée. Les réponses, rédigées simplement, s’appuient sur de grandes revues médicales et citent leurs sources. OpenEvidence est une plateforme déjà utilisée par des médecins pour retrouver rapidement des études.

Apple ne cherche pas à savoir si l’IA pose de bons diagnostics, mais comment les gens s’en servent : quand ils lui font confiance, s’ils comprennent la réponse, et ce qu’ils font ensuite (changer une habitude, consulter un médecin ou ne rien faire). Les participants notent les réponses, remplissent un questionnaire chaque mois et peuvent, s’ils le veulent, partager les données de l’app Santé (rythme cardiaque, sommeil, activité).

Côté données, Apple promet qu’elles seront anonymisées, jamais utilisées pour entraîner une IA, et effacées par OpenEvidence au bout de 24 heures. Les questions inquiétantes seront renvoyées vers un professionnel. L’étude n’est pas ouverte en France ni en Europe. Elle dit pourtant beaucoup de la suite : la vraie question n’est plus seulement ce que l’IA sait, mais à quel point on la croit.

LE DÉCRYPTAGE

Modèle de raisonnement (Reasoning model)

Aussi dit mode réflexion · thinking · IA qui réfléchit

Un modèle de raisonnement est une IA qui, avant de vous répondre, découpe le problème en étapes, les vérifie et corrige ses erreurs, au lieu d'écrire sa réponse d'un seul jet.

EN UNE IMAGE

Pensez à 2 façons de chiffrer un devis compliqué. La 1re : vous annoncez un prix de tête, tout de suite. La 2e : vous prenez une feuille de brouillon, vous listez les postes, vous faites les calculs, vous relisez, puis vous donnez le prix. Un modèle de raisonnement fait la 2e. Le brouillon, c'est lui qui le remplit, souvent sans que vous le voyiez.

Votre question → brouillon caché (étapes, calculs, vérifications) → réponse finale

ANALOGIE (POUR UN ENFANT DE 12 ANS)

En contrôle de maths, il y a l'élève qui écrit le résultat directement, et celui qui pose l'opération sur son brouillon avant de recopier. Le 2e met plus de temps, mais il se trompe beaucoup moins sur les problèmes difficiles. Pour 2 + 2, en revanche, le brouillon ne sert à rien.

Pourquoi ça compte : dans ChatGPT, Claude ou Gemini, vous voyez de plus en plus de boutons « réflexion », « raisonnement » ou « thinking ». Les activer change 3 choses. La qualité d'abord : sur un calcul, une analyse de contrat ou un problème en plusieurs étapes, les erreurs baissent nettement. Le délai ensuite : la réponse peut prendre de quelques secondes à plusieurs minutes. Le coût enfin : le brouillon est fait de jetons, et ces jetons sont facturés, même si vous ne les lisez jamais. Savoir quand activer ce mode vous évite de payer et d'attendre pour des tâches simples, ou à l'inverse d'obtenir une réponse bâclée sur une décision importante.

Exemple en entreprise : un cabinet de gestion de patrimoine de 6 personnes utilise l'IA toute la journée. Pour reformuler un mail, résumer un compte rendu ou trouver un titre de newsletter, le mode rapide suffit : réponse immédiate, coût minime. Mais quand un conseiller lui demande de comparer 3 scénarios de transmission d'entreprise, avec des abattements et des calendriers différents, il active le mode réflexion. L'IA prend 2 minutes, détaille ses calculs étape par étape, et repère une incohérence dans les chiffres fournis. Le conseiller vérifie toujours le résultat, mais il part d'un travail structuré au lieu d'une réponse approximative. La règle du cabinet : réflexion pour ce qui engage un client, mode rapide pour le reste.

À retenir : un modèle de raisonnement échange de la vitesse contre de la fiabilité : précieux sur les questions difficiles, inutile et plus cher sur les tâches simples.

À NE PAS CONFONDRE

Agent d'intelligence artificielle (AI agent) : un agent agit. Il ouvre des outils, envoie des mails, remplit des fiches. Un modèle de raisonnement réfléchit, mais ne fait rien en dehors de la conversation. Beaucoup d'agents utilisent un modèle de raisonnement comme cerveau, mais ce sont 2 choses différentes.

Recherche approfondie (deep research) : un mode qui part lire des dizaines de pages sur internet pendant plusieurs minutes pour rédiger un rapport. Le modèle de raisonnement, lui, peut réfléchir longtemps sans rien chercher : il travaille avec ce qu'il sait et ce que vous lui donnez.