

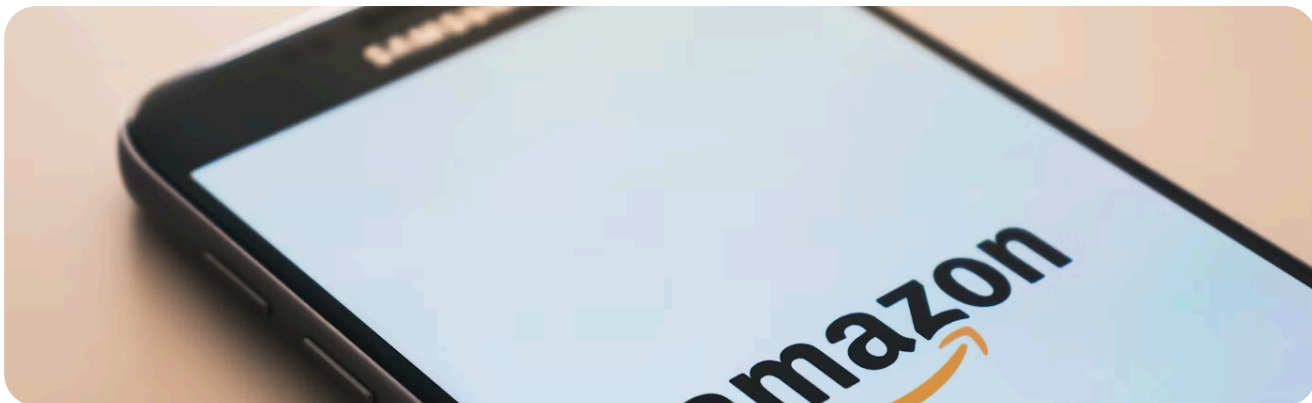
AU SOMMAIRE

5 actus · 1 Décryptage

- 01 Amazon ferme le service où des humains faisaient, à la main, le travail attribué à l'IA
- 02 Un modèle chinois réclame 30 % des revenus à Microsoft, Amazon et Google
- 03 Google installe son IA dans les cabinets d'avocats, quatre grands noms en test
- 04 Claude se souvient désormais de vous d'un outil à l'autre, sans rien demander
- 05 Samsung fait calculer la mémoire elle-même pour tripler la vitesse de l'IA

LE DÉCRYPTAGE

Agent d'intelligence artificielle (AI agent) : quand l'IA ne répond plus, elle agit.



01

Amazon ferme le service où des humains faisaient, à la main, le travail attribué à l'IA

Mechanical Turk s'arrêtera le 30 septembre, vingt et un ans après son lancement. Des centaines de milliers de personnes y étiquetaient des images et retranscrivaient des fichiers pour quelques centimes la tâche.

Ouvert en 2005, le service mettait en relation des entreprises et des travailleurs payés à la pièce pour des micro-tâches : reconnaître ce qu'il y a sur une photo, taper un enregistrement audio, répondre à un questionnaire. Jeff Bezos parlait d'« intelligence artificielle artificielle » : le client voyait un service automatique, des humains faisaient le travail derrière. Plus de 500 000 personnes y sont passées.

Amazon évoque une simple revue de ses programmes. Les inscriptions étaient déjà fermées depuis le 30 juillet, et deux services voisins d'annotation de données (le fait d'étiqueter à la main des exemples pour entraîner une IA) ont aussi été coupés aux nouveaux clients. Aucun plan de reclassement, aucune indemnité pour les travailleurs concernés : ils devront basculer vers les plateformes concurrentes montées ces dernières années pour recruter les annotateurs des grands modèles.

Le fait le plus gênant est ailleurs. Une étude universitaire de 2023 estimait que jusqu'à 46 % des travailleurs interrogés sur la plateforme utilisaient eux-mêmes des outils d'IA pour boucler leurs tâches. Autrement dit, la matière première censée être humaine était déjà en partie produite par des machines. Si vous achetez des données annotées, des tests utilisateurs ou des réponses à des questionnaires en ligne, la question à poser à votre prestataire n'est plus le prix par réponse : c'est comment il prouve que la réponse vient d'une personne.



02

Un modèle chinois réclame 30 % des revenus à Microsoft, Amazon et Google

Moonshot AI négocie pour que son modèle Kimi K3 soit proposé aux clients d’Azure, d’AWS et de Google Cloud. Les discussions viennent de commencer et peuvent encore échouer.

Kimi K3 est un modèle à poids ouverts (les fichiers du modèle sont téléchargeables, on peut donc l’installer sur ses propres machines) de 2 800 milliards de paramètres (les réglages internes appris pendant l’entraînement). C’est le plus imposant du genre disponible publiquement. Sur les problèmes qui demandent plusieurs étapes de raisonnement, des évaluateurs indépendants le placent au niveau des meilleurs modèles américains.

Le laboratoire pékinois demande jusqu’à 30 % des revenus générés par les services bâtis sur son modèle chez les trois hébergeurs. Ce serait le premier accord de ce type entre un laboratoire chinois et un géant américain du calcul à distance. Trois points bloquent encore : le partage exact, l’accès aux données des clients, et la façon de compter la consommation réelle, facturée en jetons (les fragments de texte que l’IA manipule). Un accord voisin a déjà été signé en juillet avec un acteur chinois plus modeste.

L’enjeu est concret. Si l’accord aboutit, un modèle très performant et nettement moins cher apparaît dans le catalogue de l’hébergeur que vous utilisez peut-être déjà, sans changer de fournisseur ni de facture. Reste la question qui décidera de son adoption en Europe : où passent les données, et sous quelle juridiction.



03

Google installe son IA dans les cabinets d'avocats, quatre grands noms en test

Lancée le 25 août en avant-première, une version de sa plateforme professionnelle conçue pour le droit relit les contrats, cherche la jurisprudence, surveille les changements de réglementation et prépare les dépôts au tribunal.

L'offre repose sur quatre briques : des compétences métier qui guident l'IA pas à pas dans une tâche juridique, des connecteurs vers les logiciels déjà en place dans les cabinets, des agents fournis par des éditeurs spécialisés du droit, et une couche de contrôle qui permet aux équipes informatiques et risques de voir qui accède à quoi et ce que l'IA a produit. Quatre cabinets internationaux de premier rang l'essaient déjà.

Les connecteurs visent les outils du quotidien du métier : gestion documentaire, recherche juridique, analyse de preuves, signature électronique. C'est la première déclinaison sectorielle de cette plateforme, lancée en même temps qu'une version pour les services financiers. La santé et les sciences du vivant sont annoncées ensuite.

Le droit est un secteur qui facture le temps passé. Une IA qui absorbe la relecture de contrats et la veille réglementaire attaque directement la partie la plus standardisée de cette facturation. Pour une petite structure qui achète du conseil juridique, l'effet se verra d'abord sur les prestations les plus répétitives, celles qu'un cabinet vend aujourd'hui à l'heure et qu'il pourra bientôt vendre au forfait. Et le schéma va se répéter : après le droit et la finance, chaque métier réglementé aura sa version.



04

Claude se souvient désormais de vous d'un outil à l'autre, sans rien demander

La mémoire de l'assistant est maintenant partagée entre les conversations et Cowork, son mode qui exécute des tâches à votre place. C'est activé par défaut, y compris sur l'offre gratuite.

Jusqu'ici, ce que vous expliquiez dans une discussion restait dans cette discussion. Désormais, ce que Claude retient d'un échange est disponible quand il exécute une tâche, et ce qu'il apprend en travaillant revient dans la conversation suivante. L'assistant tient un carnet vivant des sujets au fil de l'eau, au lieu d'attendre la fin d'un échange pour en faire un résumé.

Tout ce qui est retenu est visible dans une liste de fichiers et de sujets, dans les réglages. Vous pouvez la relire, corriger une entrée, en supprimer une, ou dire directement à l'assistant ce qu'il doit garder. Trois catégories restent exclues sauf accord explicite : la santé, la politique et l'identité de genre.

Le gain de temps est réel : ne plus réexpliquer un projet, un client ou une façon de rédiger à chaque nouvelle session. Le revers est qu'un dossier sur votre entreprise se constitue tout seul, dans un outil grand public, sans que personne ait eu à l'activer. Si vous travaillez sous accord de confidentialité, deux gestes suffisent à cadrer : ouvrir cette liste une fois par trimestre, et décider ce qui n'a rien à y faire.



05

Samsung fait calculer la mémoire elle-même pour tripler la vitesse de l'IA

Présentée le 25 août devant les spécialistes des puces, une mémoire d'un nouveau genre embarque une petite unité de calcul à l'intérieur. Résultat annoncé : trois fois plus de texte produit par seconde.

Aujourd'hui, faire tourner une IA revient surtout à déplacer des données entre la mémoire et le processeur, sans arrêt. C'est ce trajet, plus que la puissance de calcul brute, qui ralentit tout. L'idée de cette puce est de faire une partie des opérations directement là où les données sont stockées, pour supprimer une bonne part des allers-retours.

Les chiffres annoncés : 16 gigaoctets de capacité, et un débit interne de 614 gigaoctets par seconde, contre 76,8 gigaoctets pour une mémoire classique, soit huit fois plus. À l'usage, le fabricant mesure 3,01 fois plus de texte produit par seconde. Le point décisif est ailleurs : la puce garde le même boîtier et le même brochage que la mémoire actuelle, donc un fabricant de téléphones ou d'ordinateurs portables peut l'adopter sans redessiner sa carte électronique.

L'arrière-plan est une pénurie. Les mémoires très rapides destinées aux cartes graphiques d'IA sont accaparées par les centres de données, et leur prix tire vers le haut le coût de tous les appareils. Faire tourner l'IA sur l'appareil lui-même change deux lignes de votre budget : vos données ne partent pas chez un tiers, et il n'y a pas de facture à l'usage. C'est la voie la plus crédible pour que l'IA devienne un coût fixe plutôt qu'un abonnement qui grossit avec l'activité.

LE DÉCRYPTAGE

Agent d'intelligence artificielle (AI agent)

Aussi dit agent IA · agentique

Un agent, c'est une IA à qui vous confiez un objectif plutôt qu'une question : elle décide elle-même des étapes, se sert des outils auxquels elle a accès, et va jusqu'au résultat.

EN UNE IMAGE

Un assistant classique, c'est un GPS : vous demandez la route, il vous répond, vous conduisez. Un agent, c'est la voiture qui conduit : vous donnez l'adresse, elle choisit l'itinéraire, s'arrête au péage, repart.

Vous → objectif → l'IA choisit les étapes → mail · agenda · fichiers → résultat

ANALOGIE (POUR UN ENFANT DE 12 ANS)

Vous demandez à votre grand frère la recette des crêpes : il vous la récite, et vous faites tout. Maintenant vous lui dites « prépare le goûter ». Il ouvre le placard, voit qu'il manque du lait, descend en acheter, allume la plaque, met la table. Personne ne lui a donné la liste des étapes : il l'a écrite lui-même. Voilà un agent. Et voilà aussi pourquoi on regarde ce qu'il a acheté.

Pourquoi ça compte : le mot est partout dans les offres commerciales, et il change la question à poser. Face à un assistant, on juge la qualité de la réponse. Face à un agent, on juge les droits qu'on lui donne : quels outils, quel plafond de dépense, quelles actions il peut lancer sans validation, et quelle trace il laisse. Une erreur d'assistant produit une phrase fausse que vous corrigez. Une erreur d'agent produit une suite d'actions déjà effectuées.

Exemple en entreprise : une agence web de quatre personnes reçoit vingt demandes par semaine sur son formulaire de contact. Un agent lit les messages, vérifie dans le fichier clients si le contact existe déjà, classe la demande par type de prestation, rédige un devis à partir des tarifs de l'agence, puis s'arrête et attend une validation humaine avant tout envoi. Environ vingt minutes gagnées par demande. Le garde-fou tient en une règle : l'agent prépare, il n'écrit jamais au client sans qu'un humain ait cliqué.

À retenir : un assistant répond, un agent agit. La bonne question n'est donc plus « est-ce que la réponse est juste ? » mais « jusqu'où a-t-il le droit d'aller, et qu'est-ce qui reste tracé ? »

À NE PAS CONFONDRE

Agent conversationnel (chatbot) : il discute et il répond, mais il n'agit pas en dehors de la conversation. Il ne va pas ouvrir votre boîte mail ni modifier une fiche client.

Automatisation classique : les scénarios du type « si un formulaire est rempli, alors envoyer ce message ». Les étapes sont écrites d'avance par un humain et ne varient jamais. Un agent, lui, choisit ses étapes au moment où il agit : plus souple, et moins prévisible.