KLEAA

MARDI · 25 AOÛT 2026

AU SOMMAIRE

5 actus · 1 chiffre

01 La bibliothèque où le monde entier va chercher ses modèles d’IA est à vendre

02 Nvidia négocie une entrée au capital de Perplexity, valorisé plus de 30 milliards de dollars

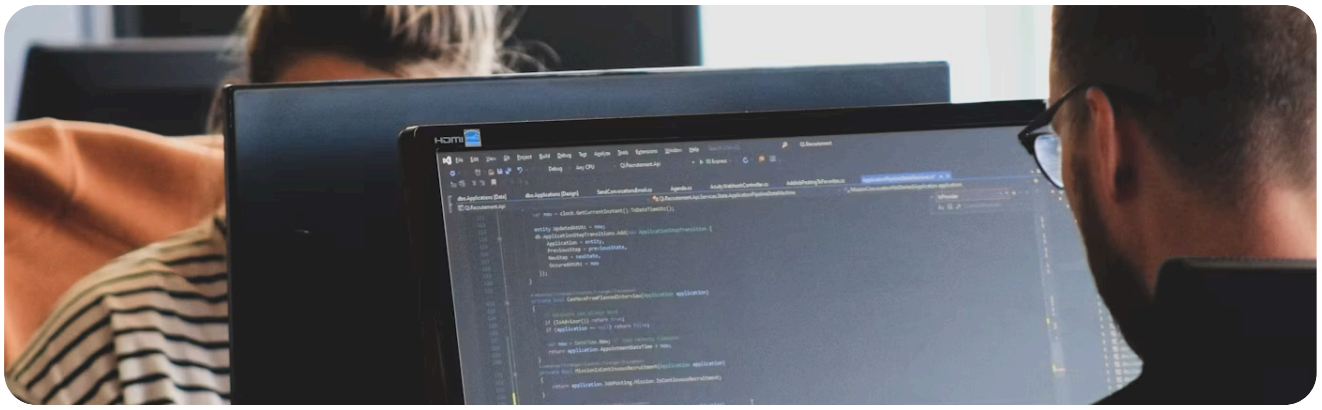
03 Un assistant IA très efficace réclame l’accès à vos mails, à votre écran et à votre micro

04 Un modèle d’IA gratuit et très performant circule sans que l’on sache qui le fabrique

05 SoftBank emprunte 6,3 milliards de dollars aux épargnants japonais pour financer OpenAI

LE CHIFFRE

40 % : part des organisations qui ont déployé l’IA à l’échelle de toute l’entreprise en 2026.



01

La bibliothèque où le monde entier va chercher ses modèles d’IA est à vendre

Hugging Face, la plateforme sur laquelle des millions de développeurs déposent et téléchargent des modèles d’intelligence artificielle, examine des offres de rachat qui la valorisent 13 milliards de dollars ou plus. Aucun acquéreur n’est identifié à ce stade.

Hugging Face n’est pas un produit grand public, et c’est pour cela que son nom vous dit sans doute peu de choses. C’est un entrepôt. Les laboratoires y publient leurs modèles, les entreprises viennent y chercher celui qui correspond à leur besoin, le testent, puis le mettent en service. Une bonne partie des outils d’IA que vous utilisez sans le savoir passe par ce catalogue. La société avait été valorisée 4,5 milliards de dollars lors de sa dernière levée de fonds, en 2023, menée par Salesforce Ventures.

Rien n’est signé. L’entreprise a mandaté une banque pour recueillir les marques d’intérêt, et les candidats ne sont pas connus. Plus tôt cette année, elle avait refusé un investissement de 500 millions de dollars de Nvidia, sur une valorisation de 7 milliards, pour ne pas dépendre d’un actionnaire dominant. Son dirigeant, Clément Delangue, affirme que la société est proche de la rentabilité et rappelle que la communauté lui confie ses données et ses modèles.

L’enjeu tient en une phrase pour qui construit un produit branché sur de l’IA. Une pièce neutre de la chaîne, aujourd’hui au service de tous les laboratoires à égalité, pourrait passer sous le contrôle de l’un d’entre eux. Le rachat d’OpenRouter par Stripe pour plus de 7 milliards de dollars allait déjà dans ce sens. Chaque brique indépendante qui change de mains ajoute une dépendance à surveiller dans un produit qu’on croyait bâti sur des fondations ouvertes.

02

Nvidia négocie une entrée au capital de Perplexity, valorisé plus de 30 milliards de dollars

Perplexity vend un moteur de recherche qui répond en phrases rédigées plutôt qu'en liste de liens. Son revenu annualisé (le chiffre d'affaires des derniers mois ramené à une année complète) est passé de moins de 250 millions de dollars en début d'année à plus de 750 millions aujourd'hui.

L'information a été publiée le 23 août par The Information. Nvidia figurerait parmi les investisseurs d'un tour de table de plusieurs milliards de dollars, sur une valorisation supérieure à 30 milliards. Ni Nvidia ni Perplexity n'ont commenté, et rien n'est bouclé. Les deux sociétés travaillent déjà ensemble, et Nvidia avait par ailleurs envisagé avec Perplexity un accord de licence de plusieurs milliards de dollars.

Le chiffre à retenir n'est pas la valorisation, c'est le revenu. Multiplier son chiffre d'affaires par trois en huit mois, sur un marché où Google reste solidement installé, signifie que des clients paient réellement pour une réponse rédigée plutôt que pour une page de résultats. Une partie de cette croissance vient de Perplexity Computer, un outil qui exécute des tâches à votre place sur l'ordinateur au lieu de se contenter de répondre à une question.

Deux choses en découlent. D'abord, la recherche par IA n'est plus une démonstration technique, c'est une activité qui encaisse. Ensuite, si vos clients prennent l'habitude de poser leur question à un assistant, votre visibilité ne dépend plus seulement du classement de votre site, mais de la façon dont ces assistants parlent de vous. Le trafic ne se mesure alors plus en clics reçus, mais en citations obtenues.

03

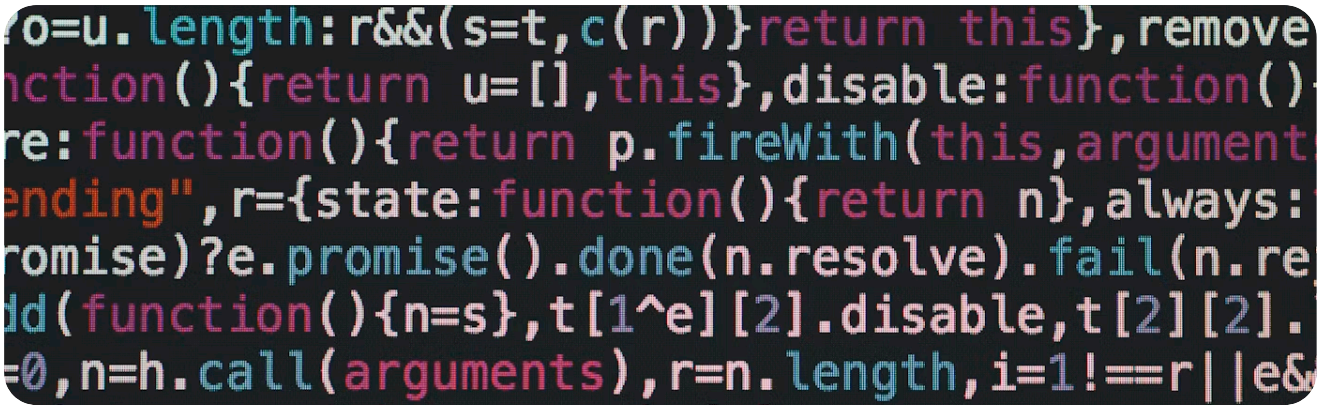
Un assistant IA très efficace réclame l’accès à vos mails, à votre écran et à votre micro

Instinct, un assistant personnel encore en accès restreint, se branche sur la messagerie, l’agenda, la localisation, l’écran et le son du téléphone pour exécuter des tâches à votre place. Ses conditions d’utilisation et ses failles alarment les premiers testeurs.

Instinct est développé par une équipe menée par Noah Shinn, ancien chercheur chez Sierra, au sein de la société Spear Street Technology, à San Francisco. On lui parle par message, par WhatsApp ou par téléphone, et il agit : il réserve, cherche des vols, achète. Pour cela il se connecte à la messagerie, aux applications de discussion, à l’agenda, et il peut lire l’écran, écouter et connaître la position de l’appareil. C’est ce qui le rend efficace, et c’est exactement ce qui pose problème.

Les conditions d’utilisation accordent à l’éditeur une licence perpétuelle et irrévocable pour accéder aux contenus de l’utilisateur, les stocker, les reproduire, les transmettre, les publier, les modifier et s’en servir pour entraîner ses modèles. Elles l’autorisent aussi à conclure des engagements au nom de l’utilisateur. Des testeurs rapportent des mails encore stockés en clair après avoir coupé l’accès, un assistant facilement trompé par un message piégé, l’extraction de codes de vérification reçus par mail sans autorisation explicite, et un mail parti sans validation préalable.

Le raisonnement vaut pour tout assistant qui agit à votre place, pas seulement celui-ci. Un outil qui dispose des droits d’accès d’un assistant de direction en porte aussi les risques : un seul compte compromis ouvre la messagerie, l’agenda, les documents et parfois les moyens de paiement. Avant de brancher un agent (un programme qui enchaîne seul plusieurs actions au lieu de répondre à une question) sur un compte professionnel, deux lectures s’imposent : la liste des autorisations demandées, et le paragraphe des conditions qui dit ce que l’éditeur fait de vos contenus.



04

Un modèle d’IA gratuit et très performant circule sans que l’on sache qui le fabrique

Depuis le 20 août, un modèle baptisé Ox Alpha est proposé gratuitement sur OpenRouter, un site qui revend l’accès à des centaines de modèles d’IA. Il obtient d’excellents résultats en programmation, et son éditeur reste anonyme.

La fiche du modèle est laconique : il est développé et exploité par un fournisseur tiers qui a choisi de rester anonyme pendant cette phase d’essai. Le modèle accepte du texte, des images et de la vidéo, traite jusqu’à un million de jetons (les unités de texte qu’un modèle lit et écrit, environ trois quarts de mot chacune) en une seule fois, et ne coûte rien. Sur DeepSWE, un test de correction de bugs réels, un développeur indépendant l’a mesuré à 80 %, devant Claude Fable 5 à 65 % et GPT-5.6 Sol à 52 %. Ce sont des mesures privées, pas des résultats officiels.

L’engouement est immédiat, le doute aussi. Le 22 août, un chercheur a comparé les traces d’erreur, les codes techniques et le découpage du texte, et conclut à une parenté avec GLM 5.3, le modèle du laboratoire chinois Zhipu AI. Aucune confirmation n’est venue. Patrick Collison, dirigeant de Stripe, a qualifié publiquement le modèle de très impressionnant. Le trafic mesuré sur OpenRouter montre déjà des milliards de jetons envoyés depuis des outils de travail comme Claude Code : ce ne sont plus des essais, c’est de la production.

La question à se poser avant d’y envoyer quoi que ce soit est simple : qui lit ? OpenRouter indique que les échanges sont conservés par le fournisseur et ne servent pas à l’entraînement. L’accès proposé par OpenCode annonce une absence totale de conservation, mais cela ne couvre que sa propre couche, pas celle du fournisseur en amont. Autrement dit, un modèle gratuit dont l’exploitant est inconnu garde une copie de vos requêtes. Pour un prototype, aucune importance. Pour du code propriétaire, un contrat ou un fichier client, la gratuité coûte cher.

05

SoftBank emprunte 6,3 milliards de dollars aux épargnants japonais pour financer OpenAI

Le groupe japonais lance la plus grosse émission d’obligations destinée aux particuliers jamais réalisée au Japon : 1 000 milliards de yens, soit 6,3 milliards de dollars, sur sept ans, pour tenir ses engagements auprès d’OpenAI.

Une obligation est un prêt : vous prêtez de l’argent à une entreprise, elle vous verse un intérêt chaque année et vous rembourse à l’échéance. Ici, les prêteurs visés sont des particuliers japonais, pas des fonds d’investissement. Le taux annoncé se situe entre 4,3 % et 4,9 %, la durée est de sept ans, et le prix doit être fixé le 4 septembre. Le précédent record d’émission auprès du grand public au Japon était déjà détenu par SoftBank.

Le détail qui compte, c’est la raison de ce montage. SoftBank s’est engagé sur plus de 60 milliards de dollars auprès d’OpenAI et cherche par ailleurs un prêt de 10 milliards garanti par sa participation dans l’entreprise. Or l’agence S&P note le groupe BB+, un cran sous la catégorie dite d’investissement. Les banques hésitent à porter seules un risque de cette taille sur sept ans, alors ce sont les épargnants qui prennent le relais.

Ce que dit ce montage sur l’état du financement de l’IA vaut d’être noté. Les sommes en jeu dépassent ce que les circuits habituels absorbent, et le besoin d’argent glisse vers la dette et vers le public. Vous n’y participez pas directement, mais vous en dépendez : les modèles que vous payez à l’usage reposent sur des engagements financiers de cette nature. Quand le coût de l’argent monte, il finit toujours par apparaître sur la facture des abonnements et des accès aux modèles.

LE CHIFFRE

40 %

des organisations ont déployé l’IA à l’échelle de toute l’entreprise en 2026

Il ne s’agit pas d’avoir testé un outil dans un coin, mais d’un déploiement généralisé : mêmes outils, mêmes accès, mêmes usages dans toutes les équipes. Ce taux était de 22 % un an plus tôt.

La nuance est décisive. Beaucoup d’entreprises déclarent utiliser l’IA dès qu’un salarié ouvre ChatGPT dans son navigateur. Cette mesure-là compte autre chose : les organisations où l’IA est installée à l’échelle de la structure entière, avec des outils choisis, des accès distribués et des usages inscrits dans le travail courant. Elles étaient 22 % en 2025, elles sont 40 % en 2026.

Un quasi-doublement en douze mois, sur un critère aussi exigeant, veut dire que le sujet a quitté les services informatiques. On ne parle plus d’une équipe qui essaie dans son coin, mais d’une décision prise en haut et appliquée partout, avec un budget, un calendrier et un responsable. C’est le moment où l’IA cesse d’être une expérimentation pour devenir une ligne de dépense comme le téléphone ou la comptabilité.

Le revers se lit tout aussi bien : six organisations sur dix n’en sont pas là. L’écart entre les deux groupes ne tient plus au budget, puisque l’accès aux modèles coûte quelques euros par mois et par personne. Il tient au fait d’avoir tranché : quels usages, par qui, avec quelles données, et qui répond du résultat quand la machine se trompe. Tant que ces questions restent sans réponse, l’IA demeure un bricolage individuel qui n’apparaît nulle part dans les comptes.