# KLEAA

JEUDI · 17 SEPTEMBRE 2026

AU SOMMAIRE

5 actus · 1 Décryptage



LE DÉCRYPTAGE

Génération augmentée par récupération (Retrieval-Augmented Generation) : l’IA qui répond avec vos documents.


01

# Une IA peut désormais installer et gérer votre WhatsApp Business

Meta a ouvert une porte d'entrée qui permet à un assistant comme Claude ou ChatGPT de créer le compte, de vérifier le numéro et d'écrire les messages types, à la place d'un développeur.

---

Annoncée lundi, la nouveauté s'appelle WhatsApp Business Tools MCP. Derrière le sigle, une idée simple : Meta publie un branchement standard (une prise commune que les assistants IA savent utiliser pour piloter un logiciel) qui donne à une IA l'accès aux réglages de WhatsApp Business. Jusqu'ici, mettre en place la messagerie professionnelle obligeait à naviguer entre la console développeur de Meta, le gestionnaire d'entreprise, la documentation technique et son éditeur de code. Désormais, on décrit ce que l'on veut, et l'assistant crée le compte, fait vérifier le numéro de téléphone, puis l'enregistre auprès du service d'envoi de Meta.

L'assistant rédige aussi les modèles de messages, ces textes pré-validés que WhatsApp impose pour une confirmation de commande ou un rappel de rendez-vous, et les corrige quand ils sont refusés. Il teste les notifications automatiques et signale en amont les blocages classiques : conditions d'utilisation non acceptées, moyen de paiement manquant, vérification d'entreprise incomplète. Ces erreurs faisaient jusqu'ici échouer une mise en service sans le moindre message d'explication.

Meta réserve pour l'instant l'outil à la mise au point, pas à l'envoi massif, et le déploie progressivement. Le mouvement reste notable pour une petite structure : la partie technique d'un canal client, celle qui coûtait deux à trois jours de prestataire, se rapproche d'une conversation. Ce qui ne disparaît pas, en revanche, c'est la validation par Meta, qui garde la main sur ce que vous avez le droit d'envoyer et à quel rythme.


02

# Google casse le prix des IA qui tiennent une conversation au téléphone

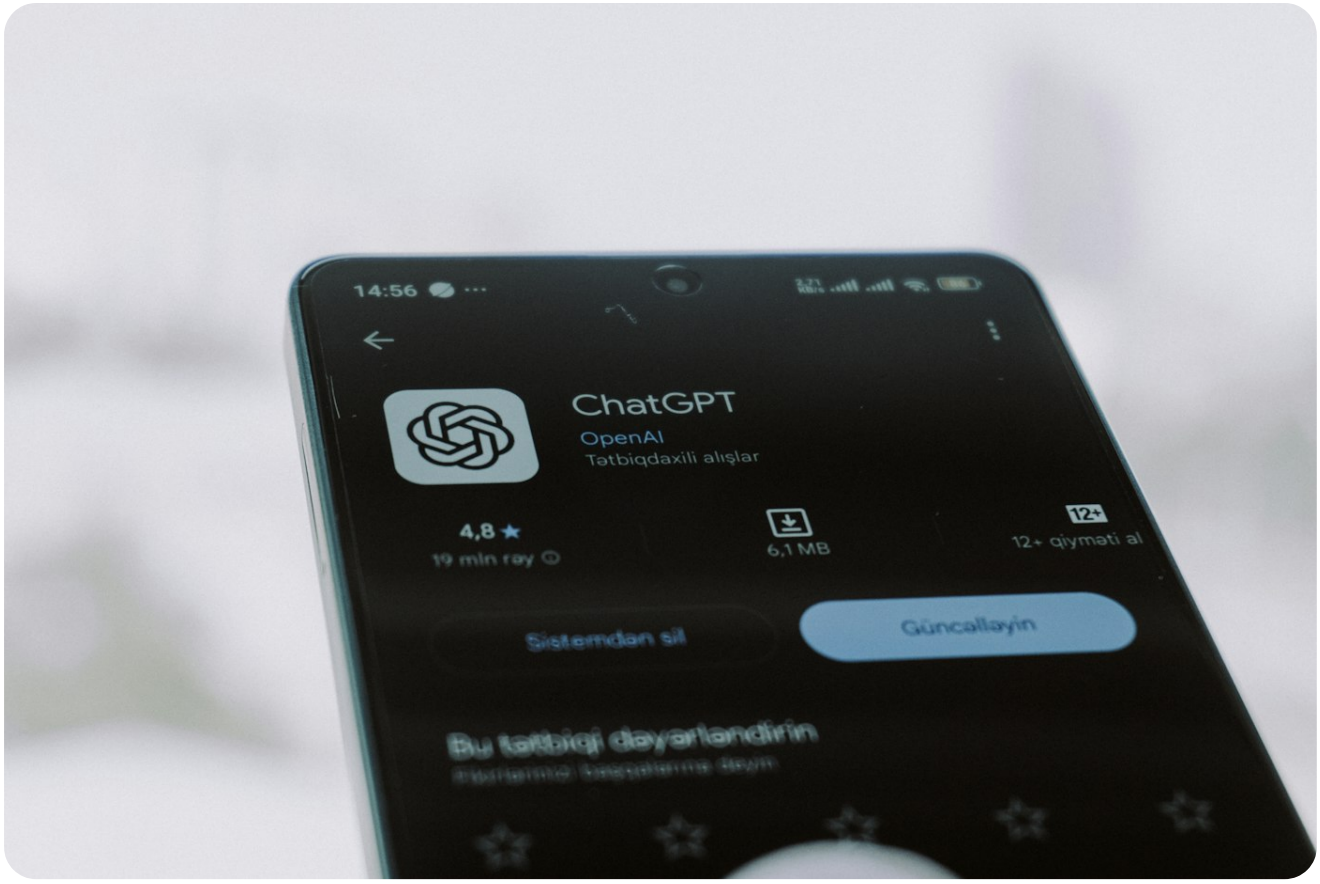
Deux nouveaux modèles vocaux sont facturés 0,005 dollar la minute écoutée et 0,018 dollar la minute parlée, soit environ moitié moins que l’offre équivalente d’OpenAI.

---

Google a présenté mardi Gemini 3.8 Live et Gemini 3.8 Live Extended Thinking. Les deux entendent et répondent directement en voix, sans repasser par une transcription écrite, ce qui supprime le délai gênant que l’on connaît sur les serveurs vocaux. Le premier vise le volume et le prix, le second les échanges compliqués qui demandent de raisonner en plusieurs étapes avant de répondre. Sur un classement indépendant de la qualité vocale, la version qui réfléchit le plus longtemps arrive première.

Le vrai sujet est le tarif. Une heure de conversation revient à environ 1,40 dollar si les deux parties parlent sans interruption, contre plus du double chez le concurrent direct. La version la plus lente et la plus fine coûte, elle, autour de 3,50 dollars l’heure. À ce niveau, la comparaison ne se fait plus entre deux fournisseurs d’IA, mais entre une IA et un centre d’appel.

Pour un artisan, un cabinet ou un commerce, cela déplace une décision très concrète : la prise de rendez-vous, la confirmation d’intervention ou le premier filtre des appels deviennent chiffrables à la minute. Restent deux garde-fous à poser avant de brancher quoi que ce soit sur votre ligne : annoncer clairement que l’interlocuteur parle à une machine, et prévoir le passage à un humain dès que la demande sort du cadre prévu.




03

# OpenAI paie des sous-traitants pour lire de vraies conversations ChatGPT

Des documents internes décrivent un programme baptisé Project Lily : des centaines de personnes recrutées pour lire les demandes des utilisateurs et noter les réponses de 1 à 7.

---

Les évaluateurs sont recrutés par une société intermédiaire, payés par une autre, et touchent plus de 50 dollars de l'heure. Leur travail consiste à juger la qualité des réponses, à résumer ce que l'utilisateur cherchait vraiment, et à signaler les défauts de ton : formulation mécanique, ou tendance à approuver tout ce que dit son interlocuteur. Ces notes servent ensuite à corriger le modèle.

Un filtre automatique est censé retirer les informations personnelles avant que la conversation n'arrive sous les yeux d'un relecteur. OpenAI reconnaît que des détails sensibles passent malgré tout. Plus délicat encore : les relecteurs voient, au-dessus de la conversation, un résumé de ce que ChatGPT a mémorisé de l'utilisateur au fil du temps, ce qui peut révéler un métier, une région ou une situation personnelle sans que rien n'ait été écrit ce jour-là.

Rien de tout cela n'est illégal, et la pratique existe chez les concurrents. Elle rappelle surtout une règle de base : une conversation avec un assistant grand public n'est pas un espace privé. Deux réflexes valent d'être pris cette semaine. Aller couper, dans les réglages, l'option qui autorise l'utilisation de vos échanges pour améliorer les modèles. Et rappeler à vos équipes qu'un contrat, un fichier de paie ou un dossier client n'a rien à faire dans une fenêtre de discussion ouverte avec un compte personnel.


04

# Bruxelles veut fermer les chatbots aux moins de 15 ans

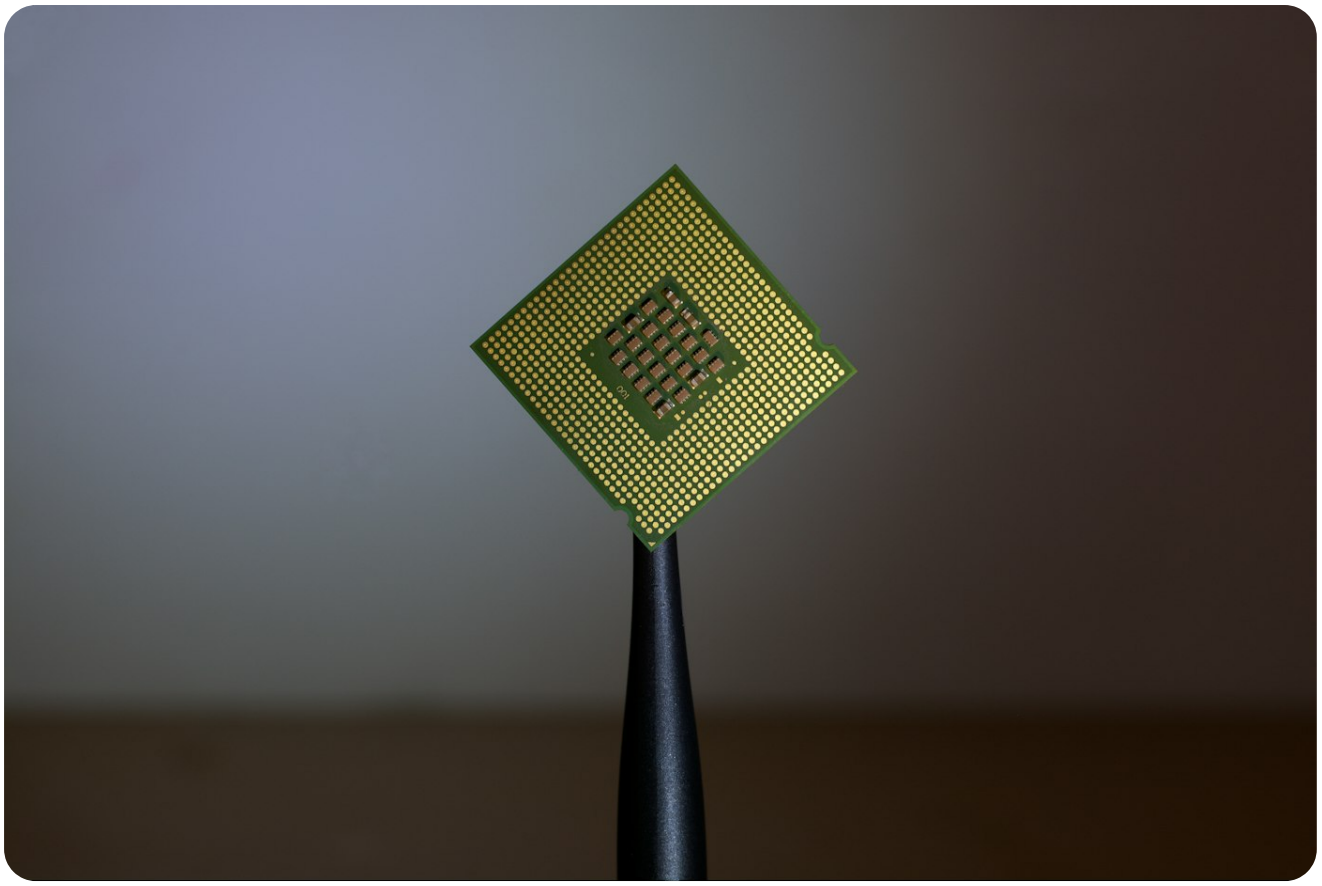
Dans son discours sur l’état de l’Union, Ursula von der Leyen a proposé d’interdire aux enfants l’accès aux réseaux sociaux, aux jeux en ligne et aux assistants conversationnels, avec vérification de l’âge à l’inscription.

---

Le principe annoncé hier à Strasbourg tient en trois paliers. Avant 13 ans, aucun accès aux réseaux sociaux, aux jeux en ligne ni aux chatbots (ces assistants avec lesquels on discute en écrivant). Entre 13 et 15 ans, pas de compte personnel, mais un compte encadré par un parent ou un tuteur. Au-delà, l’accès redevient normal. La vérification de l’âge se ferait à la création du compte, et avant le téléchargement pour les plateformes de jeux.

Le texte prévoit aussi une redevance versée par les plateformes pour financer les contrôles, c’est-à-dire faire payer la surveillance par ceux qui sont surveillés. La proposition doit être présentée dans la foulée, puis négociée par les États membres et le Parlement, ce qui prendra des mois. Elle intervient alors que plusieurs pays européens avançaient déjà seuls sur l’âge minimum.

L’effet dépasse largement les grandes plateformes. Toute entreprise qui met un assistant conversationnel en libre accès sur son site, sans compte ni identification, entre potentiellement dans le champ du texte. Deux questions méritent d’être posées dès maintenant à qui fournit votre outil : sait-il distinguer un mineur d’un adulte, et que se passe-t-il concrètement le jour où la réponse doit être oui.



05

# Une jeune pousse néerlandaise lève 200 millions d’euros pour des puces moins gourmandes

Euclyd, fondée il y a deux ans près d’Eindhoven, a bouclé un tour de table mené par Samsung. L’ancien patron d’ASML en prend la présidence.

---

La levée, annoncée mardi, dépasse 200 millions d’euros, soit environ 231 millions de dollars, pour une société créée en 2024 et qui n’a encore rien vendu. Elle est co-menée par Samsung, aux côtés de Somerset Capital Partners, du fonds Scaleup Europe géré par EQT et d’Innovation Industries. Peter Wennink, qui a dirigé ASML, le fabricant néerlandais des machines sans lesquelles aucune puce moderne ne sort d’usine, devient président du conseil.

Euclyd ne vise pas l’entraînement des modèles, cette phase très coûteuse où une IA apprend. Elle vise l’inférence, le moment où un modèle déjà entraîné répond à une question. C’est là que se joue la facture quotidienne d’une entreprise qui utilise l’IA, et c’est le terrain que Nvidia domine aujourd’hui. La promesse tient en une phrase : la même réponse pour beaucoup moins d’électricité. Les premiers systèmes ne sont pas attendus avant 2028.

Deux lectures pour un dirigeant. D’abord, l’Europe finance enfin la partie matérielle et pas seulement les logiciels, avec un argent asiatique et un savoir-faire local. Ensuite, ce genre d’annonce rappelle où va le coût de l’IA : pas dans l’abonnement mensuel, mais dans l’électricité consommée à chaque réponse. Ce poste-là finira par apparaître dans vos factures de fournisseurs.

LE DÉCRYPTAGE

# Génération augmentée par récupération (Retrieval-Augmented Generation)

Aussi dit RAG · l'IA branchée sur vos propres documents

C'est le fait de donner à une IA vos documents au moment où vous posez la question, pour qu'elle réponde à partir de ces documents-là plutôt que de sa seule mémoire.

### EN UNE IMAGE

Un expert et son armoire à dossiers. Sans ce mécanisme, vous interrogez quelqu'un qui répond de tête, avec ce qu'il a lu il y a deux ans. Avec, un assistant part d'abord chercher les 3 bonnes pages dans votre armoire, les pose sur la table, et l'expert répond en les ayant sous les yeux.

```
Votre question → recherche dans vos documents → extraits → IA → réponse
```

### ANALOGIE (POUR UN ENFANT DE 12 ANS)

Une interrogation à livre ouvert. C'est le même élève, avec la même tête, mais il a le droit d'ouvrir le manuel avant de répondre. Il se trompe beaucoup moins, et surtout il peut montrer la page d'où vient sa réponse. Si le manuel est périmé, en revanche, il recopiera sagement l'erreur.

**Pourquoi ça compte** : un assistant grand public connaît ce qui traînait sur internet jusqu'à une certaine date. Il ignore vos tarifs, vos délais, vos conditions de garantie et vos procédures internes. Ce mécanisme est le moyen le plus rapide de combler ce trou sans réentraîner quoi que ce soit : mettre à jour revient à remplacer un fichier. Il apporte aussi deux choses qui manquent cruellement ailleurs, la possibilité d'exiger que chaque réponse cite le passage dont elle sort, et celle de retirer un document pour qu'il cesse d'être utilisé. Devant un devis, trois questions suffisent : où sont stockés mes documents, qui peut les lire, et la réponse indique-t-elle sa source.

**Exemple en entreprise** : une entreprise de 12 personnes qui installe des pompes à chaleur vit avec 400 pages de notices, une trentaine de grilles tarifaires et les conditions de garantie de 5 fabricants. Avec ce mécanisme, un technicien en clientèle demande quelle pression de service s'applique à tel modèle encore sous garantie, et reçoit la réponse accompagnée des 2 pages exactes. Quand un fabricant change sa notice, on remplace le fichier et c'est terminé. Là où ça déraille : si deux versions du tarif traînent encore dans le dossier partagé, l'outil citera l'une des deux au hasard. Le vrai chantier n'est pas l'IA, c'est le ménage dans vos documents.

**À retenir** : ce mécanisme ne rend pas l'IA plus intelligente, il lui met les bons documents sous les yeux au bon moment.

### À NE PAS CONFONDRE

**Réglage fin (fine-tuning)** : on réentraîne le modèle sur des exemples pour changer sa façon de répondre, son ton, son format. Cela modifie ce qu'il est, pas ce qu'il a sous les yeux. Pour des informations qui bougent, comme un tarif ou un stock, c'est la récupération de documents qu'il faut, pas le réentraînement.

**Moteur de recherche interne** : il vous rend une liste de fichiers et vous laisse lire. Ici, la recherche n'est qu'une étape : l'IA lit les extraits à votre place et rédige une réponse, ce qui fait gagner du temps mais rend la vérification de la source indispensable.