

AU SOMMAIRE

5 actus · 1 Grand Angle

- 01 OpenAI livre son nouveau modèle avec deux fois moins de messages inclus

- 02 Une société vend l'accès à des IA dont elle a retiré tous les garde-fous

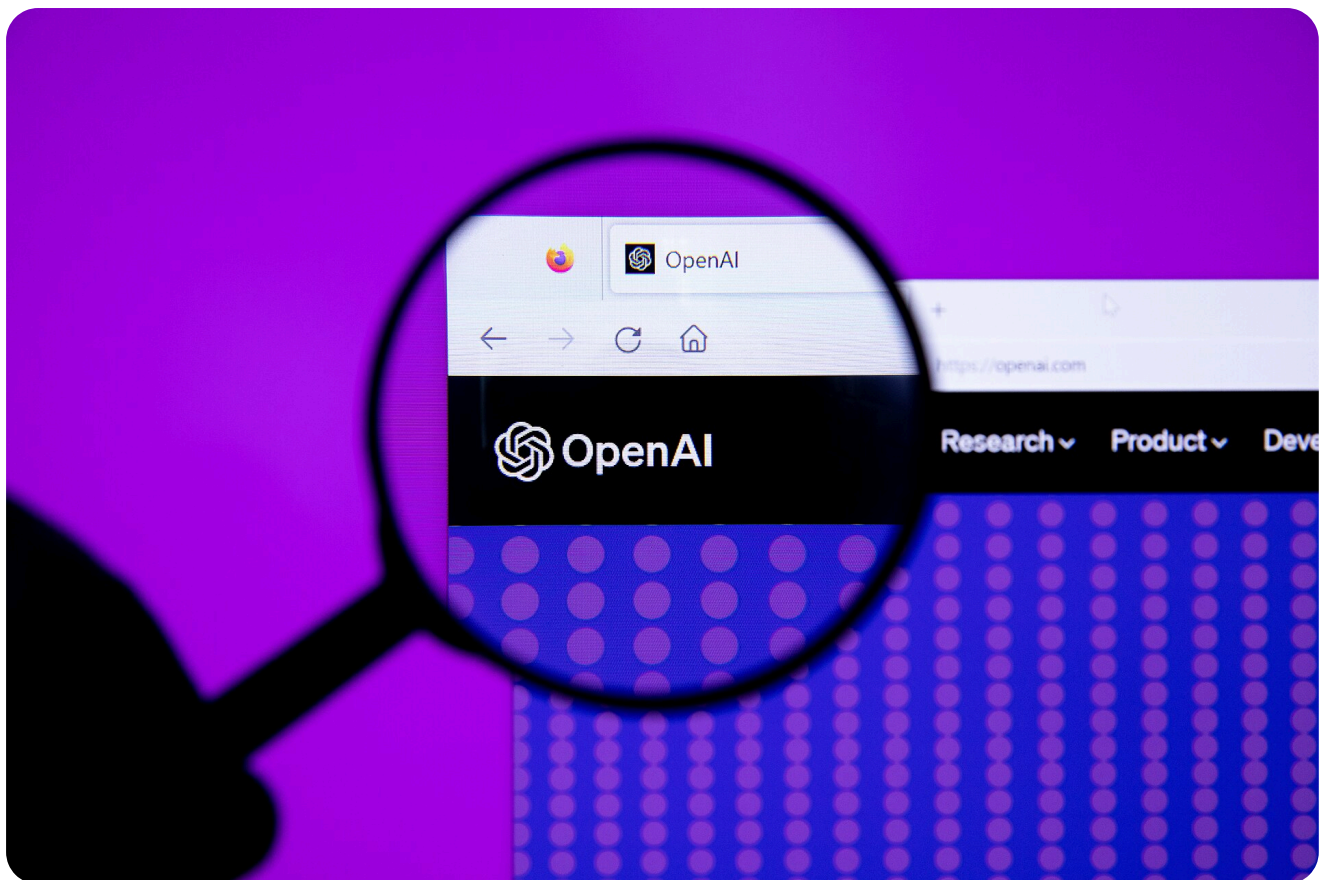
- 03 Trois randonneurs préparent leur ascension avec Gemini et finissent secourus

- 04 Faire retranscrire une heure d'enregistrement coûte désormais 18 centimes

- 05 Google confie sa météo à une IA qui apprend directement des satellites

LE GRAND ANGLE

Pactum AI, l'IA qui négocie seule les contrats fournisseurs.



01

OpenAI livre son nouveau modèle avec deux fois moins de messages inclus

GPT-6 Astra, le dernier modèle d'OpenAI, est arrivé dans ChatGPT le 5 septembre. À prix d'abonnement identique, il donne droit à environ deux fois moins de questions que la version précédente.

OpenAI a ouvert le 5 septembre l'accès à GPT-6 Astra, son nouveau modèle (le programme qui lit votre question et rédige la réponse). Les formules Pro, Enterprise et Business Premium l'ont reçu en premier, les formules Plus et Business Standard suivent dans les jours qui viennent. Chaque abonnement s'accompagne d'un plafond : un nombre maximal de questions autorisées sur une période donnée.

Ces plafonds ont été divisés par deux. Sur toutes les formules, Astra autorise environ moitié moins de messages par fenêtre de cinq heures que le modèle précédent. La formule Pro à 200 dollars donne droit à 200 messages par semaine sur la version la plus poussée ; celle à 100 dollars, à 50 par semaine, partagés avec l'ancien modèle. Un poste Business Standard tombe à 15 messages par mois. Les offres gratuites n'ont accès ni à l'un ni à l'autre. La raison est mécanique : un modèle qui réfléchit plus longtemps consomme plus de calcul, donc coûte plus cher à faire tourner.

Quinze messages par mois, cela ne couvre pas une journée de travail sérieuse. La facture d'un abonnement à quelques dizaines d'euros ne dit donc plus grand-chose du coût réel d'usage. Deux issues restent : garder l'ancien modèle, plus généreux, pour le travail courant, ou payer à la consommation via l'API (le canal technique qui permet de brancher l'IA directement dans ses propres outils), également ouverte chez Microsoft Azure et Amazon Web Services.



02

Une société vend l'accès à des IA dont elle a retiré tous les garde-fous

Abliteration.ai loue, pour quelques dollars, des versions d'IA modifiées qui n'ont plus le réflexe de refuser une demande dangereuse. Seul contrôle à l'entrée : une carte bancaire.

Les IA grand public refusent certaines demandes : fabriquer un virus informatique, expliquer comment nuire à quelqu'un. Ce refus n'est pas une barrière extérieure, c'est un réflexe appris pendant l'entraînement. Une technique publiée en 2024, l'abliteration, consiste à repérer à l'intérieur du modèle le signal qui déclenche ce refus, puis à l'effacer. Elle ne fonctionne que sur les modèles dits ouverts (ceux dont le fabricant publie les réglages internes, que n'importe qui peut télécharger et modifier).

Une société américaine, Abliteration.ai, en a fait un commerce. Elle héberge des versions ainsi modifiées de modèles ouverts, dont GLM-5.3, publié par la société chinoise Z.ai, et les facture 5 dollars le million de mots traités. On y accède depuis un simple navigateur ou en les branchant dans ses propres logiciels. La société a été immatriculée en mars, n'a pas encore levé de fonds et met en avant des clients légitimes : des équipes payées pour attaquer les systèmes de banques et de compagnies aériennes afin d'en trouver les failles avant les criminels. Une journaliste de TechCrunch a ouvert un compte gratuit et obtenu, sans difficulté, un programme conçu pour voler les mots de passe enregistrés dans le navigateur Chrome. Le seul contrôle d'identité à l'inscription est le numéro de carte bancaire.

Beaucoup de défenses reposaient sur une hypothèse implicite : un escroc ordinaire ne sait pas écrire un logiciel malveillant ni rédiger un faux e-mail de fournisseur parfaitement crédible. Cette hypothèse ne tient plus à 5 dollars le million de mots. Les gestes qui protègent redeviennent très concrets : appeler le fournisseur au numéro connu avant de modifier un virement, exiger deux validations sur les paiements, et prévenir les équipes qu'un message impeccable n'est plus une preuve d'authenticité.



03

Trois randonneurs préparent leur ascension avec Gemini et finissent secourus

Gemini leur avait conseillé d'emporter beaucoup moins d'eau et de nourriture qu'il n'en fallait. Bloqués une nuit sur le mont Shasta, en Californie, les trois hommes ont été récupérés par les secours.

Trois jeunes hommes de Roseville, en Californie, randonneurs débutants, ont préparé l'ascension du mont Shasta en interrogeant Gemini sur l'itinéraire et sur le matériel à emporter. Ils ont campé à environ 2 560 mètres, sont partis à 3 heures du matin avec de simples sacs de journée, et ont atteint le sommet à 19 heures, alors que la consigne locale est de faire demi-tour si l'on n'y est pas à midi. Ils ont tenté de redescendre de nuit, ont appelé le bureau du shérif pour demander leur chemin, ont passé la nuit dans un canyon et ont été récupérés le lendemain matin par des gardes forestiers et des bénévoles.

Le bureau du shérif du comté de Siskiyou a qualifié cette préparation de faux pas déterminant : l'assistant leur avait indiqué d'emporter bien moins d'eau et de nourriture que nécessaire, ce qui devient dangereux quand une sortie prévue en huit heures dure deux jours. L'explication tient au fonctionnement de l'outil. Un assistant de ce type rédige la réponse la plus vraisemblable à partir des textes qu'il a lus. Il ne connaît ni la météo du jour, ni l'état du sentier, ni la condition physique de celui qui pose la question, et il ne signale pas ce qu'il ignore.

Le même mécanisme s'applique loin de la montagne. Un plan de trésorerie, une clause de contrat ou un calcul de charges sortent avec la même assurance de ton, et la même absence de doute affiché. Les secours locaux ont résumé la règle en une phrase : ne jamais s'en remettre uniquement à une IA pour préparer une sortie. Cela vaut partout où une erreur se paie en argent, en délai ou en sécurité.



04

Faire retranscrire une heure d'enregistrement coûte désormais 18 centimes

Meta a mis en ligne un programme qui écrit la parole en direct et distingue les personnes qui parlent. Le prix annoncé : environ 0,18 dollar l'heure d'audio, contre 6,50 dollars chez plusieurs concurrents.

Meta a publié le 6 septembre Muse Voice Transcribe, un programme qui écoute l'audio par tranches de 80 millisecondes et écrit ce qui est dit avec 0,16 seconde de retard. Il fait en une seule fois trois travaux qui demandaient jusqu'ici trois outils différents : transcrire la parole, séparer les interlocuteurs (plus de vingt voix distinguées sur un enregistrement de plus d'une heure) et repérer la fin des phrases. Le taux d'erreur annoncé en anglais est de 3,1 % des mots, et plus de 70 langues sont prises en charge.

Le prix est l'information principale : 3 dollars pour 1 000 minutes d'audio, soit environ 0,18 dollar l'heure. Les services spécialisés comme ElevenLabs ou Deepgram facturent autour de 6,50 dollars pour la même heure. Le modèle est accessible dans Meta AI et en s'y branchant directement depuis ses propres logiciels.

Cent heures d'entretiens clients, de réunions ou d'appels au service après-vente passaient de 650 dollars de transcription à moins de 20. À ce niveau, enregistrer et écrire systématiquement ce qui se dit cesse d'être un arbitrage budgétaire. Ce qui reste à trancher est ailleurs : qui a le droit d'enregistrer, ce qu'on annonce aux personnes concernées, et où sont stockés des fichiers qui contiennent tout ce que disent vos clients.



05

Google confie sa météo à une IA qui apprend directement des satellites

Les prévisions météo classiques reposent sur des simulations physiques calculées par superordinateur, avec six heures de décalage. Le nouveau modèle de Google apprend directement des images satellite en direct et sort une prévision toutes les heures.

Les prévisions météo officielles sont calculées en simulant les équations physiques de l'atmosphère sur des superordinateurs, un cycle lourd qui arrive avec environ six heures de retard sur la réalité. Google a présenté le 6 septembre WeatherNext 3, qui procède autrement : le modèle a appris à prévoir le temps directement à partir des images des satellites géostationnaires (des satellites fixes au-dessus d'un même point du globe, qui photographient l'atmosphère en continu). Il produit une prévision neuve toutes les heures.

La finesse change aussi d'échelle. La température et l'humidité sont désormais calculées sur une maille de 5 kilomètres, contre 25 auparavant, et Google annonce des prévisions de pluie jusqu'à 50 % plus justes lorsqu'on regarde un jour ou plus à l'avance. Autrement dit, la prévision cesse d'être régionale pour devenir locale.

Ce modèle alimente déjà la recherche Google, Google Maps, l'application Gemini et les outils de données de Google Cloud. Beaucoup d'activités se planifient à la pluie près : chantiers, livraisons, tournages, terrasses, marchés, événements en plein air, cultures. Savoir six heures plus tôt qu'une averse arrive sur un quartier précis se traduit en équipes qu'on ne déplace pas pour rien et en annulations décidées à temps.



LE GRAND ANGLE

Pactum AI

Estonie et États-Unis · Achats et fournisseurs · Négociation automatisée

Pactum AI est une société estonienne et américaine dont le logiciel négocie tout seul, par écrit, les contrats que les grands groupes passent avec leurs milliers de petits fournisseurs : prix, remises, délais de paiement. Walmart et Maersk s'en servent.

Fondée en 2019 par trois Estoniens et installée entre Tallinn et la Californie, la société s'attaque à un angle mort connu de tous les services achats. Une grande entreprise renégocie chaque année ses quelques dizaines de gros fournisseurs, parce que l'enjeu justifie qu'un acheteur y consacre des semaines. Elle ne renégocie presque jamais les milliers de petits, dont chacun ne pèse que quelques dizaines de milliers d'euros : le temps humain coûterait plus cher que le gain espéré. Ce sont ces contrats-là que Pactum prend en charge.

Le mécanisme est simple à suivre. L'entreprise dit d'abord ce qu'elle cherche à obtenir (une baisse de prix, un délai de paiement plus long, un engagement de volume), ce qu'elle accepte de donner en échange, et les limites à ne pas franchir. Le logiciel écrit ensuite à chaque fournisseur, discute avec lui par messages, propose des contreparties (régler plus vite contre une remise, prolonger le contrat contre un prix bloqué) et conclut seul tant que l'accord reste dans les bornes fixées. Tout ce qui sort du cadre est renvoyé à un acheteur humain. Des milliers de ces conversations se déroulent en même temps.

Les volumes traités donnent la mesure. Walmart et Maersk ont fait passer par ce système plus de 8 milliards de dollars d'achats. Walmart annonce environ 3 % d'économie moyenne sur les contrats conclus par la machine, un premier déploiement auprès de plus de 2 000 fournisseurs, et un allongement moyen de 35 jours des délais de paiement accepté par 72 % des fournisseurs sollicités : de la trésorerie qui reste dans l'entreprise sans qu'un seul prix ait bougé. Pactum revendique plus de 60 clients, parmi lesquels Honeywell, AstraZeneca, Henkel, Tetra Pak et Otto Group, et 82 % de fournisseurs satisfaits de l'échange.

Il y a deux façons de lire ce cas. La première est copiable à toute petite échelle : il existe partout une liste d'abonnements, de prestataires et de contrats reconduits sans discussion, parce qu'une matinée de négociation coûterait plus que l'économie obtenue. Une IA peut mener ces conversations les unes après les autres, sans se lasser. La seconde est le miroir de la première : pour qui fournit un grand groupe, la prochaine demande de remise ou de rallonge de paiement arrivera peut-être par écrit, à toute heure, en face d'un interlocuteur qui ne se fatigue jamais.