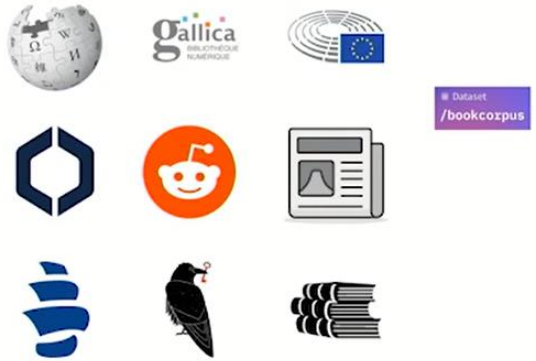
(Mandatory?) AI Commons

Jean Cattan

10/12/2025

Classer les sources de données d'entraînement

	Accéder gratuitement à ces données est-il légal ?	Utiliser ces données pour l'entraînement est-il légal ?
Données blanches	Oui	Oui
Données grises	Oui	Pas clair / Non
Données noires	Non	Non



Benoît Sagot

Directeur de Recherche Inria en Traitement Automatique des Langues et Linguistique Informatique
Responsable de l'équipe-projet ALMAnaCH
Titulaire d'une chaire dans l'institut PRAIRIE
Titulaire de la chaire annuelle « informatique et sciences numériques » du Collège de France (2023-2024)
Membre élu de la Commission d'Évaluation d'Inria

Sur quelles données les grands modèles sont-ils entraînés ?

- Parfois des données blanches et grises uniquement
  - OSCAR 2019 pour CamemBERT, ROOTS pour BLOOM
- Le plus souvent (probablement) sur des données blanches, grises et noires
  - Très peu de modèles récents – même libres ! – **décrivent** leurs données de pré-entraînement
    - Exemple : LLaMA 1
  - Encore moins les rendent **téléchargeables**
    - Reproduction approchée du corpus d'entraînement de LLaMA 1 : RedPajama

	Training Data
Llama 1	*See Touvron et al. (2023)*
Llama 2	*A new mix of publicly available online data*

(Touvron et al. 2023b)

TECHNOLOGY

Search LibGen, the Pirated-Books Database That Meta Used to Train AI

Millions of books and scientific papers are captured in the collection's current iteration.

By Alex Reisner

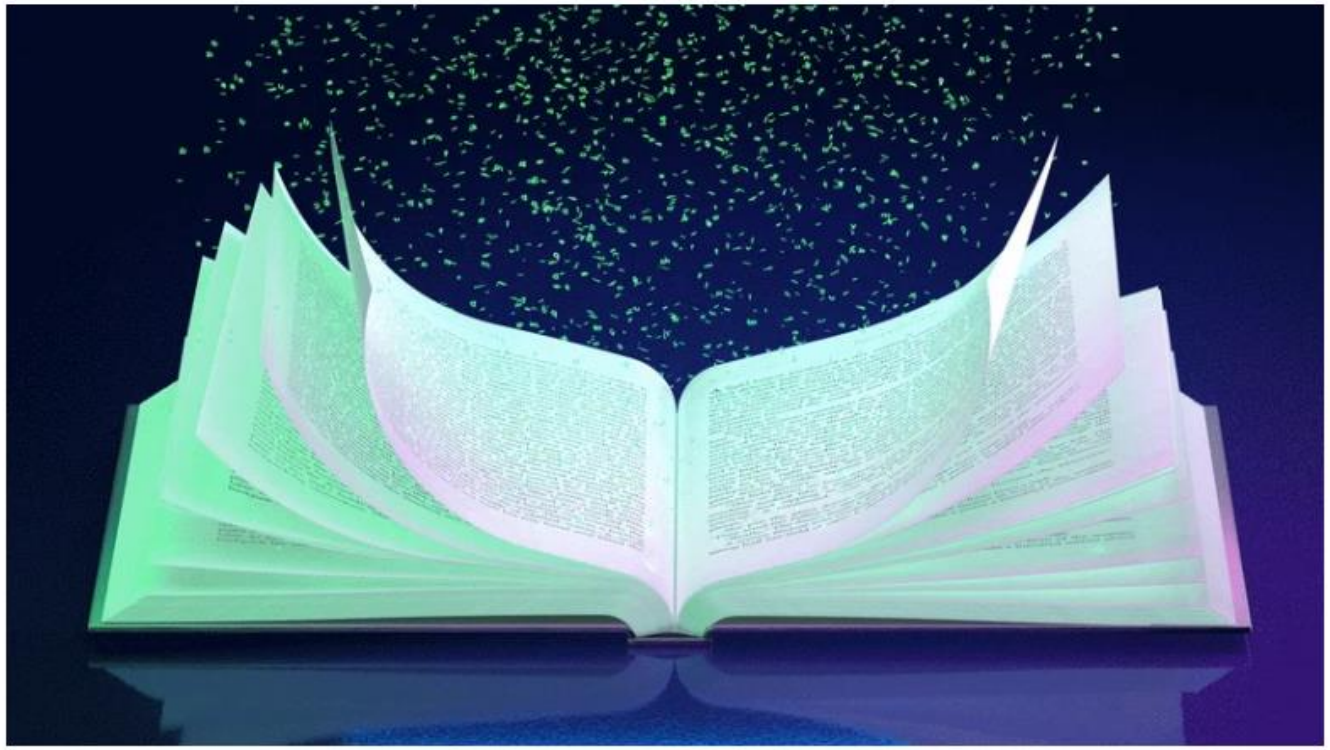


Illustration by Matteo Giuseppe Pani / The Atlantic

PRESS RELEASE | Dec 9, 2025 | Brussels | 3 min read

Commission opens investigation into possible anticompetitive conduct by Google in the use of online content for AI purposes

The European Commission has opened a formal antitrust investigation to assess whether **Google** has breached EU competition rules by **using the content of web publishers**, as well as **content uploaded on** the online video-sharing platform **YouTube, for artificial intelligence ('AI') purposes.** The investigation will notably examine whether Google is distorting competition by imposing unfair terms and conditions on publishers and content creators, or by granting itself privileged access to such content, thereby placing developers of rival AI models at a disadvantage.

Takeaways

- Meta AI will now offer a broader range of real-time content, including global news, entertainment, lifestyle stories, and more across our apps and devices.
- As a first step, we are partnering with publications such as CNN, Fox News, Le Monde Group, People Inc., and USA TODAY Co., Inc.
- Our goal over time is to provide something for everyone by continuing to add new content sources and topics.

We're beginning to offer a wider variety of real-time content on Meta AI — from global, breaking news to entertainment, lifestyle stories, and more. When you ask Meta AI news-related questions, you'll now receive information and links that draw from more diverse content sources to help you discover timely and relevant content tailored to your interests.

These integrations will also facilitate easier access to information by linking out to articles, allowing you to visit these partners' websites for more details while providing value to partners, enabling them to reach new audiences.

As the first step in our content expansion, we're partnering with a variety of outlets — CNN, Fox News, Fox Sports, Le Monde Group, the People Inc. portfolio of media brands, The Daily Caller, The Washington Examiner, USA TODAY and the USA TODAY Network. We'll continue to add new partnerships and explore new features to enhance the experience for the people who use our products.

cybersécurité : comment mieux se protéger, notamment face aux *advanced persistent threats*, etc.

Que se passe-t-il dans les pays où la communauté est plus petite ?

R.G. : La résilience d'une version de Wikipédia se mesure à la taille de sa communauté. Plus il y a de contributeurs actifs, moins il y a de chances que l'encyclopédie se retrouve noyautée par un groupe idéologique, par un groupe politique ou quoi que ce soit. L'exemple le plus marquant, c'est celui de Wikipédia en croate, qui avait une petite communauté. Un petit groupe qui s'assumait totalement comme néo-nazi a fini par faire peur aux contributeurs, qui ont fui. Ce groupe a alors pris le contrôle total de l'encyclopédie, ce qui a obligé la Fondation à fermer toute l'encyclopédie dans cette langue pour la relancer de manière plus accueillante.

On sait que Wikipédia sert aujourd'hui de base d'entraînement, au niveau des données, pour tous les grands modèles de langage. Depuis le déploiement de ces IA il y a environ trois ans, quelles ont été les principales conséquences que vous avez observées ?

R.G. : La principale conséquence, c'est une baisse du nombre de lecteurs qu'on estime autour de 8 %, et qui reflète les changements dans la manière d'accéder à l'information : aujourd'hui, beaucoup de gens s'informent directement *via* leur LLM, et ces derniers ne font pas ou peu référence à Wikipédia, ou alors cela dépend de la manière dont on formule la requête. Cela diminue mécaniquement la visibilité de l'encyclopédie. Cette décroissance pourrait poser problème sur le long terme. Moins de lecteurs signifie potentiellement moins de donateurs, donc moins de ressources pour héberger l'encyclopédie, assurer les mises à jour techniques nécessaires, mais aussi moins de moyens pour soutenir le développement des communautés. Plus une communauté est affaiblie, plus le contenu de l'encyclopédie devient fragile ou moins fiable. Étant donné que les IA continueront de s'entraîner sur ce contenu, on peut imaginer un phénomène de dégénérescence en cascade. Ce n'est pas encore le cas, mais c'est clairement un scénario que l'on a en tête et que l'on étudie.

JEUDI 11 DÉCEMBRE 2025
83e ANNÉE – N° 25177
3,80 € – FRANCE MÉTROPOLITAINE
WWW.LEMONDE.FR –
FONDATEUR : HUBERT BEUVE-MÉRY
DIRECTEUR : JÉRÔME FENOGLIO

Le Monde

Budget de la « Sécu » : pari réussi pour Lecornu

▸ A une courte majorité, l'Assemblée nationale est parvenue, mardi, à adopter le projet de loi de financement de la Sécurité sociale pour 2026

▸ Les députés Renaissance, MoDem, PS et ceux du groupe LIOT ont largement approuvé ce texte face à l'opposition massive des élus RN et « insoumis »

▸ « Cette majorité de responsabilité montre que le compromis n'est pas un slogan : il permet d'avancer dans le sens de l'intérêt général », a salué M. Lecornu

▸ Une fois ce projet de loi définitivement voté, le premier ministre s'attaquera à un tout autre défi : faire adopter le budget de l'Etat

PAGES 8-9

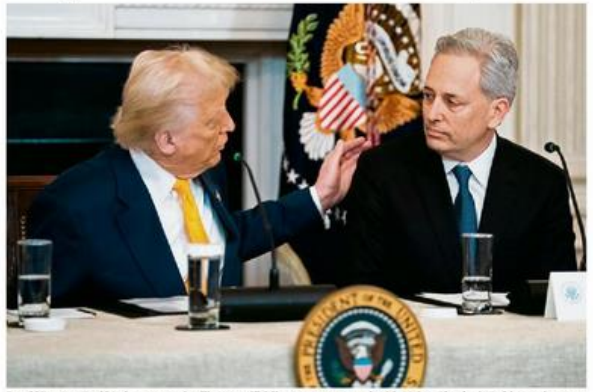
ÉDITORIAL

L'APPRENTISSAGE DOULOUREUX DU COMPROMIS

PAGE 31

DAVID SACKS, LE « MONSIEUR IA » DE TRUMP

▸ Il est le fer de lance des patrons de la Silicon Valley à la Maison Blanche
▸ Une enquête du « New York Times » l'accuse de conflits d'intérêts multiples PAGE 15



Donald Trump et David Sacks, expert en intelligence artificielle et en cryptomonnaies, à la Maison Blanche, à Washington, le 7 mars.
ALLISON ROBBERT/« THE WASHINGTON POST »/VIA GETTY

Ukraine Le dilemme militaire et moral de Zelensky sur le Donbass

ALORS QUE LE « PLAN de paix » présenté par les Etats-Unis prévoit que l'Ukraine retire ses troupes de la partie qu'elle contrôle toujours dans le Donbass, en échange de vagues garanties de sécurité, Volodymyr Zelensky a réaffirmé, lundi, son opposition à cette proposition. Le président ukrainien estime qu'il n'a « aucun droit légal » ni « moral » de céder des territoires où habitent environ 200 000 personnes, et où des centaines de milliers de soldats ont combattu depuis le début de l'invasion russe.

Dans les régions occupées d'Ukraine, un récent rapport de l'ONG Eastern Human Rights Group et de l'Institut de recherche stratégique et de sécurité documente la façon dont Moscou transforme le système éducatif pour inculquer aux enfants ukrainiens que la Russie est leur patrie.

PAGE 5

Trafic de drogue Opération « Trident » : l'itinéraire hors norme d'un indic

La justice tente de cerner le rôle d'« Elias », informateur de haut niveau des services de police, soupçonné de trafic de cocaïne

ENQUÊTE PAGES 12-13

Reportage Au Chili, les électeurs de Parisi en faiseurs de rois

Les électeurs du candidat populiste, arrivé troisième au premier tour de la présidentielle, sont très convoités

PAGE 6

Fait divers La « pensée vide » du jeune Paul D., quintuple meurtrier

PAGE 14

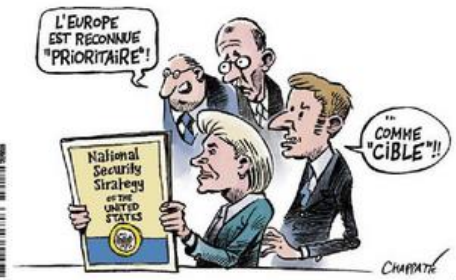
Transport pétrolier Ces nouveaux pavillons qui cachent la flotte fantôme russe

PAGE 16

Grèce Ellinikon, « mini-Dubaï » sur la Riviera athénienne

PAGE 19

VU PAR CHAPPATTE (SUISSE) — CARTOONING FOR PEACE



Algérie Le cercle des chroniqueurs désunis et désenchantés

Une bande de six auteurs prodiges et impertinents avait émergé sous la présidence Bouteflika. La répression qui a suivi le Hirak a ruiné leurs espoirs

PAGES 22-23

Arts Escher, maître fascinant des jeux d'optique

L'artiste néerlandais, « mathémagicien » de la perspective, fait l'objet d'une première grande exposition à la Monnaie de Paris

PAGE 24

Allemagne 4,20 €, Andorre 4,20 €, Autriche 4,20 €, Belgique 4,20 €, Espagne 4,20 €, Grèce 4,20 €, Guadeloupe-Martinique 4,20 €, Italie 4,20 €, Luxembourg 4,20 €, Maroc 32 DH, Pays-Bas 4,20 €, Portugal cont. 4,20 €, La Réunion 4,20 €, Sénégal 2 500 F CFA, Suisse 4,80 CHF, Tunisie 7,10 DT

Communiqué de presse

LA FRANCE, L'ALLEMAGNE, LES PAYS-BAS ET L'ITALIE CRÉENT UN CONSORTIUM POUR LES COMMUNS NUMÉRIQUES

Retrouvons notre puissance numérique : la France, l'Allemagne, les Pays-Bas et l'Italie créent un consortium pour les communs numériques

Publié le mercredi 29 octobre 2025 | DINUM

Thématiques : COMMUNS NUMÉRIQUES

La Commission européenne a approuvé la création de l'**EDIC Digital Commons (European Digital Infrastructure Consortium)**, un **nouveau cadre européen** qui permettra aux États membres de **concevoir, déployer et gérer ensemble des infrastructures numériques transfrontalières**, dotées d'une gouvernance partagée et d'une personnalité juridique propre.