Université de Lille

Réalités des usages des publications dans les IA généralistes

Madeleine Géroudet

A quelles questions peut-on essayer de répondre ?

Adrien Moyaux, Schéma d'une boîte noire : entrées, sorties et boîte au contenu inconnu, 2018

- Les IAG utilisent-elles les publications scientifiques ?
- Où les IAG viennent-elles prendre les publications scientifiques ?
- De quelle manière les IAG utilisent-elles les publications scientifiques ?
- Comment les acteurs de la diffusion des publications scientifiques réagissent-ils ?

A quelles questions peut-on essayer de répondre ?

Adrien Moyaux, Schéma d'une boîte noire : entrées, sorties et boîte au contenu inconnu, 2018

- **Les IAG utilisent-elles les publications scientifiques ? OUI !**
- Où les IAG viennent-elles prendre les publications scientifiques ?
- De quelle manière les IAG utilisent-elles les publications scientifiques ?
- Comment les acteurs de la diffusion des publications scientifiques réagissent-ils ?

Pour répondre à une question, une IAG peut utiliser…

- **Son modèle de langage, fondé sur ses données d'entraînement**, selon des mécanismes de mémorisation qui demeurent encore mal connus.

- **De sources d'information** interrogées en temps réel (web, corpus de publications…)

Explorer, aspirer, piller ?
Les publications scientifiques dans les IAG

Où viennent puiser les IAG ?

Sur le web librement accessible (sous droits ou non)
Corpus C4 de Google

Dans des sources piratées
Corpus Books1, 2 et 3

Dans des sources fermées, acquises par des voies légales
Partenariat OpenAI / Le Monde

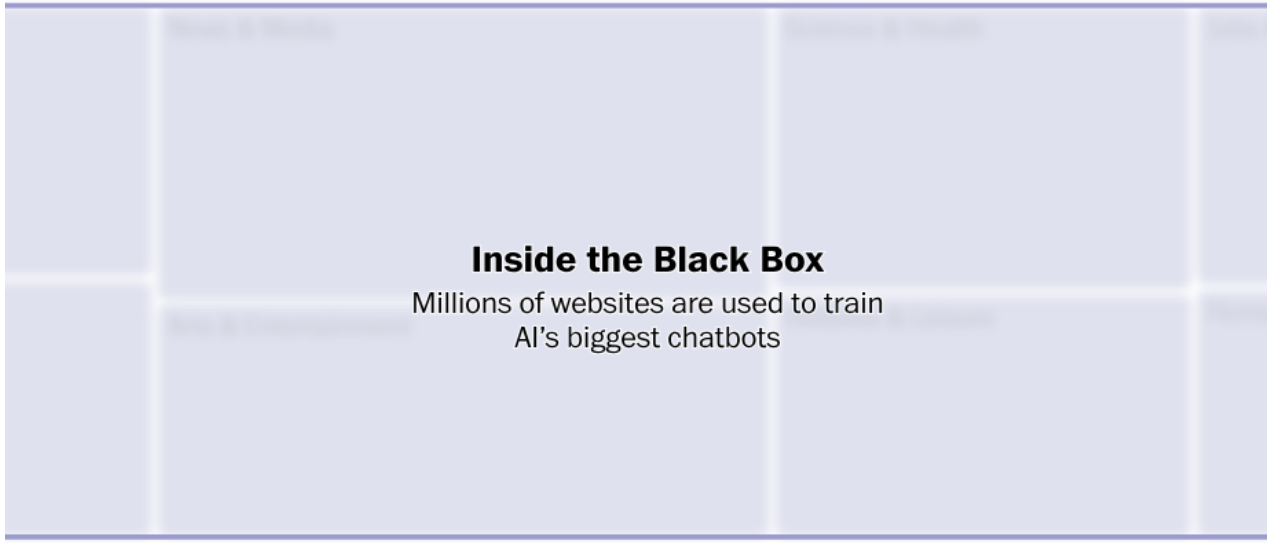
Dans des sources ouvertes
Common Corpus de PleIAs

La boîte noire, entrouverte

- Analyse du corpus de données C4 de Google, par le Washington Post et Allen Institute for IA, en 2023
- Offrant une possibilité de chercher dans la liste des sites utilisés
- Présente l'intérêt d'analyser la composition d'un corpus utilisé pour les IA sous différents angles

The Washington Post
Democracy Dies in Darkness

Tech companies have grown secretive about what they feed the AI. So The Washington Post set out to analyze one of these data sets to fully reveal the types of proprietary, personal, and often offensive websites that go into an AI's training data.



To look inside this black box, we analyzed Google's C4 data set, a massive snapshot of the contents of 15 million websites that have been used to instruct some high-profile English-language AIs, called large language models, including Google's T5 and Facebook's LLaMA.

https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Quelques exemples

Université de Lille

The websites in Google's C4 dataset

Search for a website: openedition.org

1 domain begins with "openedition.org"

RANK	DOMAIN	TOKENS	PERCENT OF ALL TOKENS
3,372,713	openedition.org	4.8k	0.000003%

The websites in Google's C4 dataset

Search for a website: theses.fr

1 domain begins with "theses.fr"

RANK	DOMAIN	TOKENS	PERCENT OF ALL TOKENS
325,235	theses.fr	72k	0.00005%

The Post believes it is important to present the complete contents of the data fed into AI models, which promise to govern many aspects of modern life. Some websites in this data set contain highly offensive language and we have attempted to mask these words. Objectionable content may remain.

Note: Some websites were unable to to be categorized and, in many cases, are no longer accessible.

The websites in Google's C4 dataset

Search for a website: arxiv

18 domains begin with "arxiv"

RANK	DOMAIN	TOKENS	PERCENT OF ALL TOKENS
2,036	arxiv.org	3.4M	0.002%

The websites in Google's C4 dataset

Search for a website: hal

11,417 domains begin with "hal"

RANK	DOMAIN	TOKENS	PERCENT OF ALL TOKENS
1,511	halfbakery.com	4.0M	0.003%
3,926	hallwynne.com	2.3M	0.001%
4,291	halfbakedharvest.com	2.1M	0.001%
4,355	hal.archives-ouvertes.fr	2.1M	0.001%
6,936	halifaxlive.com	1.5M	0.001%
8,039	halifaxcourier.co.uk	1.4M	0.0009%
8,548	hal.inria.fr	1.3M	0.0008%
8,631	halalfocus.net	1.3M	0.0008%
9,179	halowaypoint.com	1.2M	0.0008%
9,842	halopedia.org	1.2M	0.0007%

and 11,407 more results

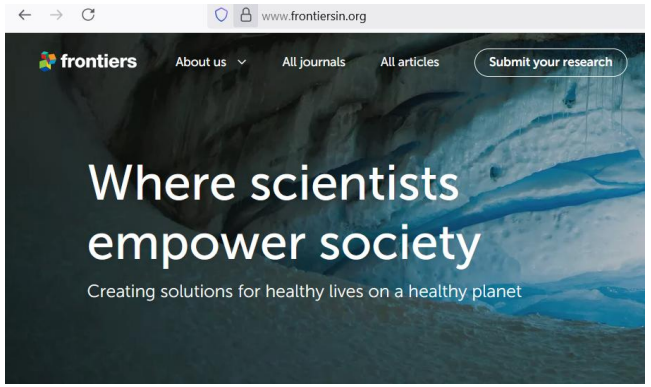
Et dans le top 15

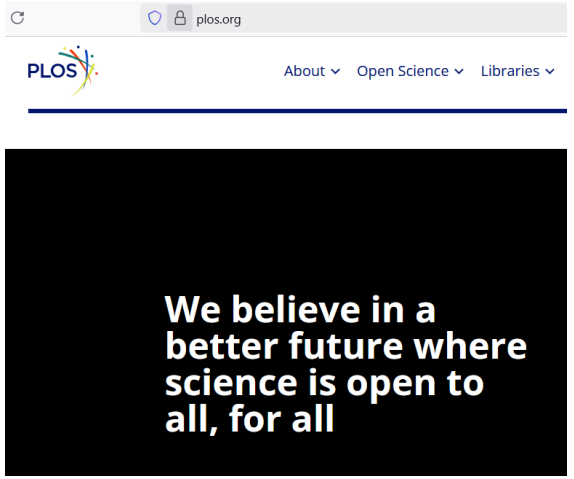
The websites in Google's C4 dataset

Search for a website

Page 1 of 67

RANK	DOMAIN	CATEGORY	PERCENT OF ALL TOKENS
1	patents.google.com	Law & Government	0.46%
2	wikipedia.org	News & Media	0.19%
3	scribd.com	News & Media	0.07%
4	nytimes.com	News & Media	0.06%
5	journals.plos.org	Science & Health	0.06%
6	latimes.com	News & Media	0.05%
7	theguardian.com	News & Media	0.05%
8	forbes.com	News & Media	0.05%
9	huffpost.com	News & Media	0.04%
10	patents.com	Law & Government	0.04%
11	washingtonpost.com	News & Media	0.03%
12	coursera.org	Jobs & Education	0.03%
13	fool.com	Business & Industrial	0.03%
14	frontiersin.org	Science & Health	0.03%
15	instructables.com	Technology	0.03%





Ouvrez la boîte noire, les contenus se referment

Depuis la diffusion des contenus du C4 :

- 5% des contenus du corpus ont été entièrement **bloqués.**
- 45% des contenus font l'objet de **restrictions d'accès.**

Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee et al. « Common Corpus : the Largest Collection of Ethical Data for LLM Pre-training. » juin 2025

Les données d'entraînement sont-elles mémorisées par les IAG ?

- Des chercheurs sont parvenus à faire citer des livres à des IAG, avec des procédés plus ou moins complexes
- Des protections anti-copie empêchent les IAG de citer facilement des textes, en particulier sous droits
- Les résultats les plus probants ont été obtenus avec Harry Potter, probablement pour des raisons de doublons dans leur corpus entraîné

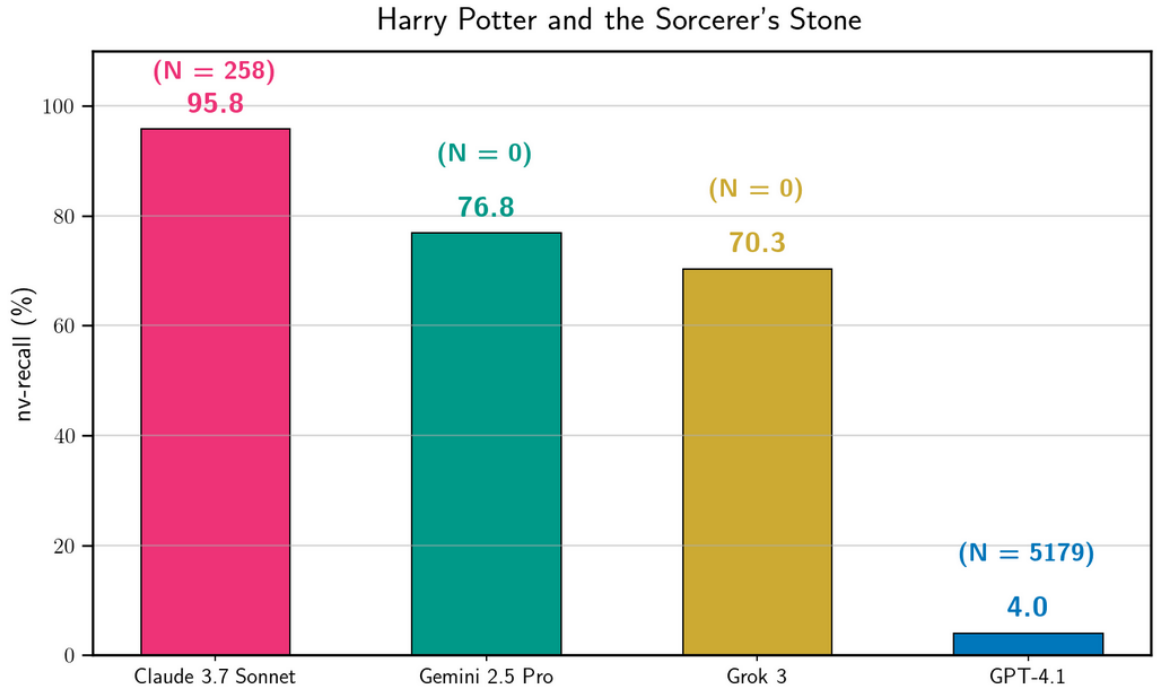


Figure 1: **Extraction of *Harry Potter and the Sorcerer's Stone* for a single run.** We quantify the proportion of the ground-truth

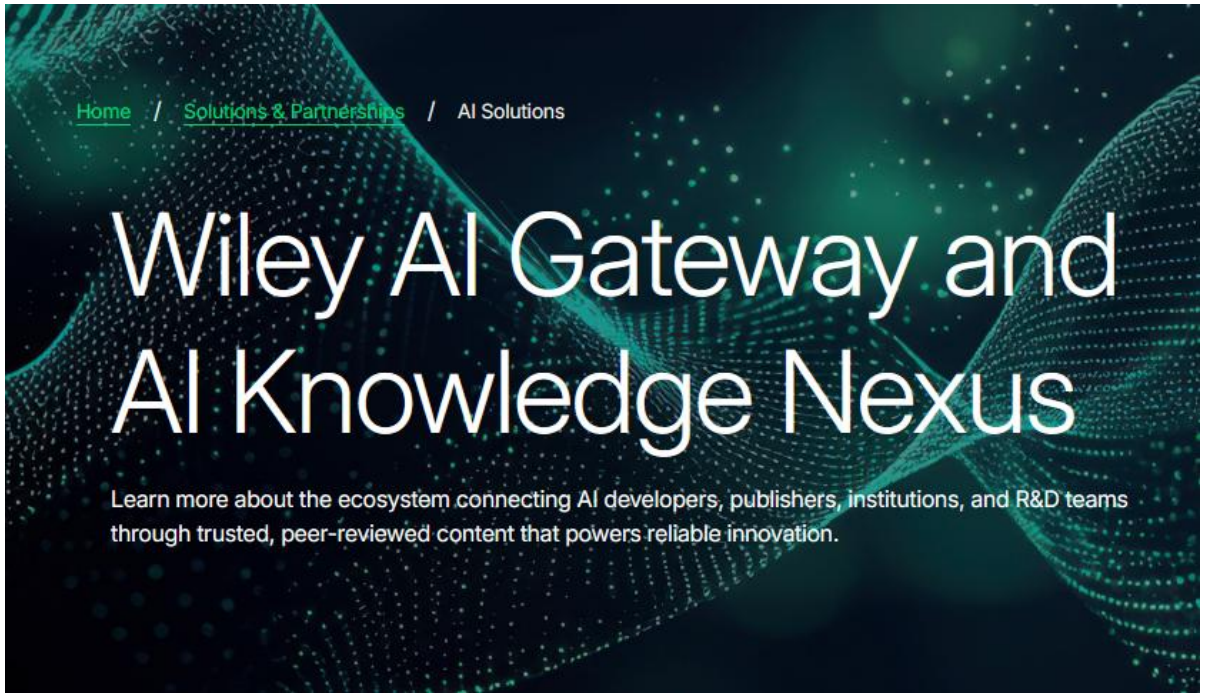
« Extracting books from production language models », Ahmed Ahmed, A. Feder Cooper, Sanmi Koyejo, Percy Liang

https://arxiv.org/html/2601.02671v1

« Comment des chercheurs ont réussi à faire citer à des IA des livres soumis au droit d'auteur », Nicolas Six, *Le Monde*

https://www.lemonde.fr/pixels/article/2026/01/22/comment-des-chercheurs-ont-reussi-a-faire-citer-a-des-ia-des-livres-soumis-au-droit-d-auteur_6663719_4408996.html

Un exemple de partenariat légal, dans le domaine de l'édition scientifique



https://www.wiley.com/en-fr/solutions-partnerships/ai-solutions/

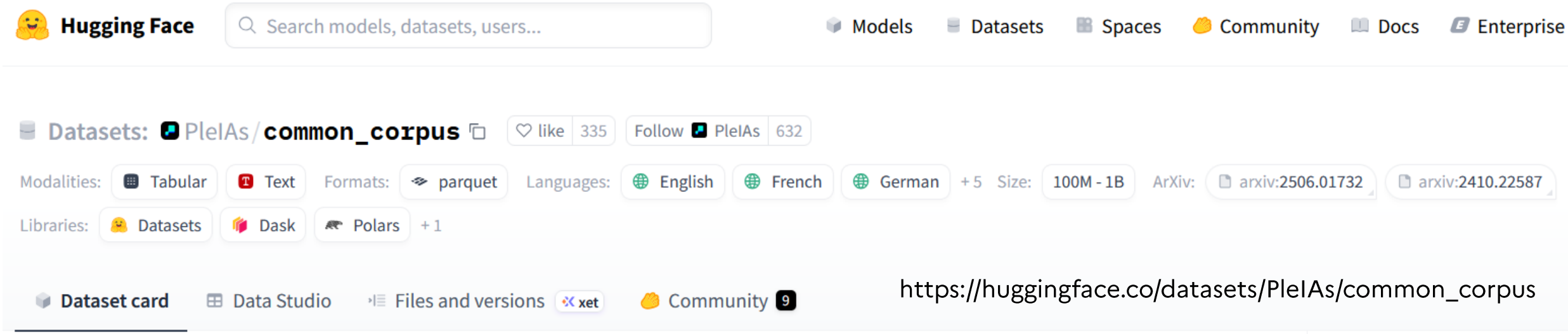
- **AI Gateway**

Mise à disposition des contenus de Wiley pour des IA généralistes (Claude, Perplexity, Mistral…)

- **AI Knowledge Nexus**

Mise à disposition de corpus pour l'entraînement de modèles

Un exemple de corpus créé à partir de sources ouvertes : Common corpus de PleIAs



https://huggingface.co/datasets/PleIAs/common_corpus

6 collections, dont :

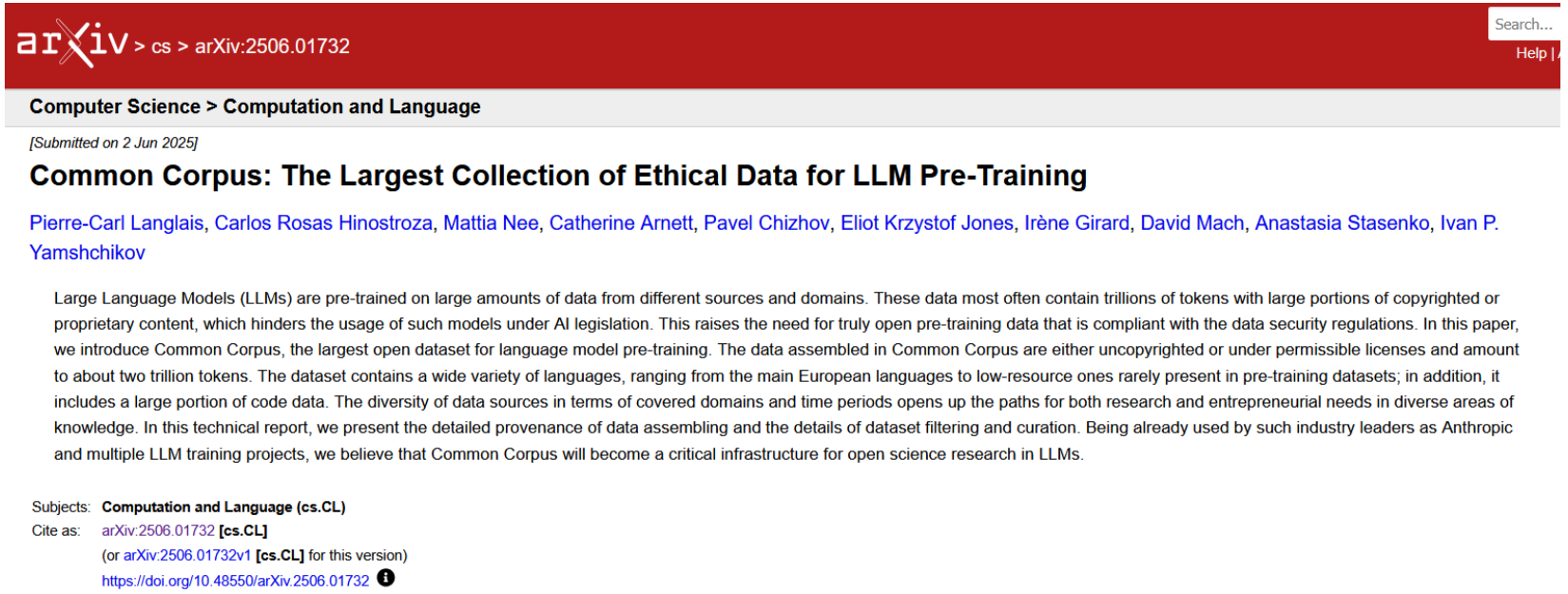
- **OpenScience :** Open Alex et autres entrepôts de science ouverte
- **OpenWeb :** contenus de Wikipédia
- **OpenCulture :** patrimoine numérique et numérisé

Dataset	Documents	Words	Tokens
Open Government	74,727,536	257,233,670,261	406,581,454,455
Open Culture	93,156,602	549,608,763,966	885,982,490,090
Open Science	19,220,942	147,305,783,453	281,193,563,789
Open Code	202,765,051	77,669,169,092	283,227,402,898
Open Web	96,165,348	33,208,509,065	73,217,485,489
Open Semantic	30,072,707	23,284,201,782	67,958,671,827
Other	925,462	328,160,421	486,099,734
Total	517,033,648	1,088,638,258,040	1,998,647,168,282

Table 1: Dataset composition of Common Corpus. For each collection, we report the total number of documents, words (whitespace-separated), and tokens.

https://arxiv.org/abs/2506.01732

Focus sur la collection Open Science du Common Corpus



arXiv > cs > arXiv:2506.01732

Search... Help | A

Computer Science > Computation and Language

[Submitted on 2 Jun 2025]

Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training

Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, Pavel Chizhov, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, Ivan P. Yamshchikov

Large Language Models (LLMs) are pre-trained on large amounts of data from different sources and domains. These data most often contain trillions of tokens with large portions of copyrighted or proprietary content, which hinders the usage of such models under AI legislation. This raises the need for truly open pre-training data that is compliant with the data security regulations. In this paper, we introduce Common Corpus, the largest open dataset for language model pre-training. The data assembled in Common Corpus are either uncopyrighted or under permissible licenses and amount to about two trillion tokens. The dataset contains a wide variety of languages, ranging from the main European languages to low-resource ones rarely present in pre-training datasets; in addition, it includes a large portion of code data. The diversity of data sources in terms of covered domains and time periods opens up the paths for both research and entrepreneurial needs in diverse areas of knowledge. In this technical report, we present the detailed provenance of data assembling and the details of dataset filtering and curation. Being already used by such industry leaders as Anthropic and multiple LLM training projects, we believe that Common Corpus will become a critical infrastructure for open science research in LLMs.

Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:2506.01732 **[cs.CL]**
(or arXiv:2506.01732v1 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2506.01732

- Des collections ouvertes, mais difficilement exploitables
- Un usage central d'Open Alex pour identifier les licences et textes intégraux
- Un corpus où l'anglais est majoritaire

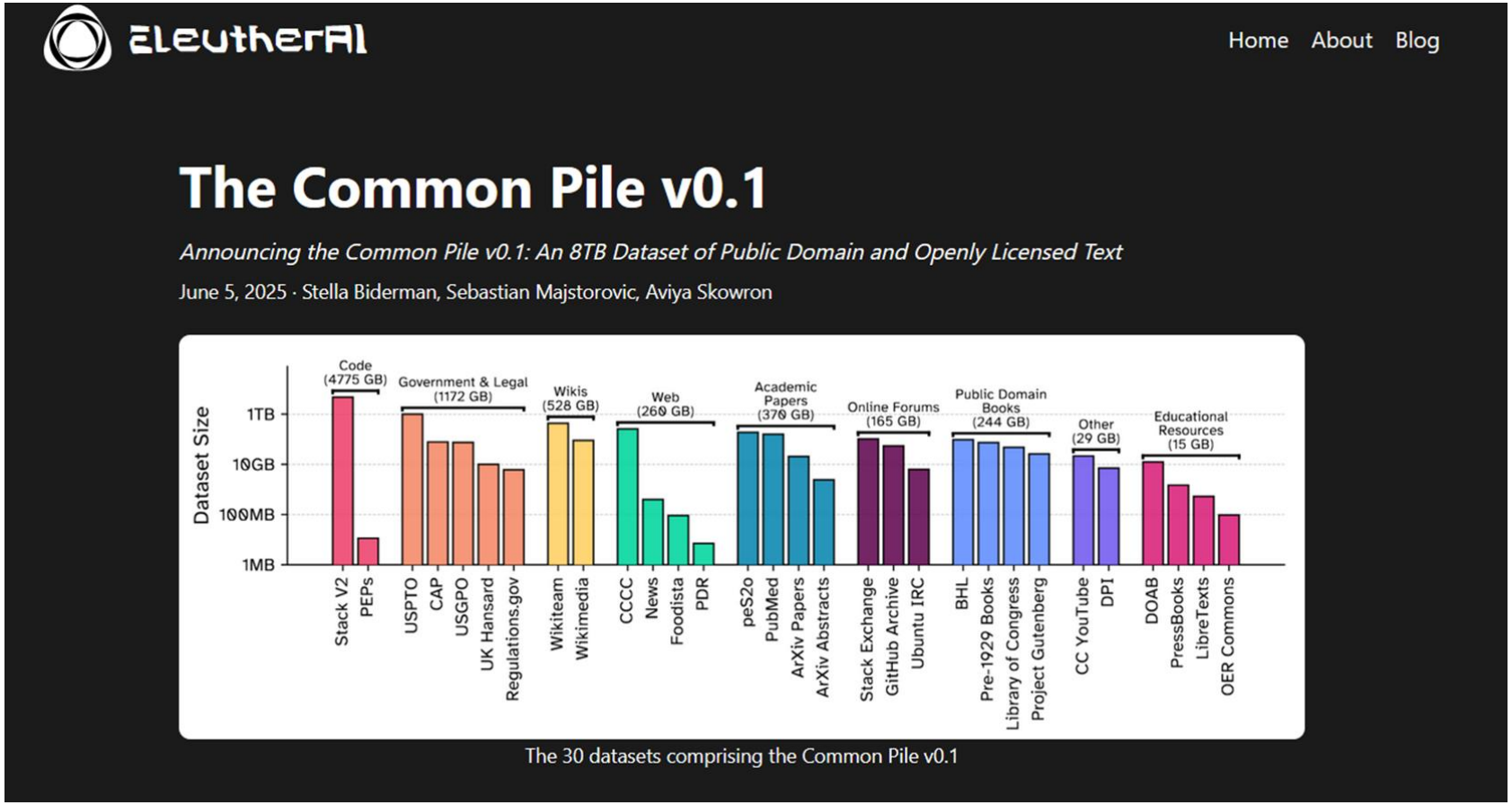
"Following the Open Source Initiative's definition of open-source AI, our data is open in terms of openness of use, meaning that use is permitted for "any purpose and without having to ask for permission" *(Open Source Initiative, 2024)*."

Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee et al. « Common Corpus : the Largest Collection of Ethical Data for LLM Pre-training. » juin 2025

License type	Tokens
Public Domain	**1,138,508,375,958**
CC-By	287,749,264,457
MIT	142,694,227,607
CC-By-SA	74,768,060,836
Apache-2.0	68,750,977,037
BSD-3-Clause	18,483,944,333
Open license	10,432,513,767
BSD-2-Clause	5,497,145,480
CC-BY-4.0	2,110,966,243
CC0-1.0	1,877,206,195

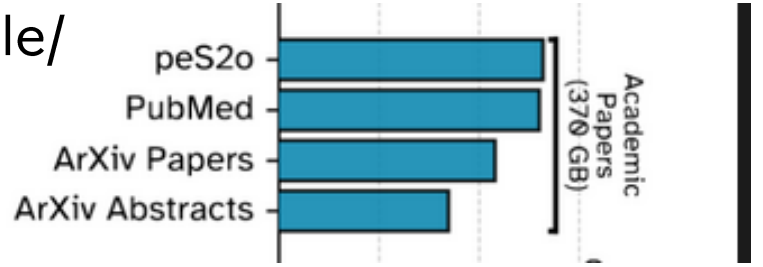
Table 3: Token counts for the ten most common licenses in Common Corpus.

Une démarche évolutive



https://blog.eleuther.ai/common-pile/

- Eleuther.ai, institut de recherche à but non lucratif
- Met à disposition The Pile en 2020, corpus de données massivement utilisé par les IAG, malgré des contenus sous droits
- En 2025, met à disposition The Common Pile, en indiquant « avoir appris de ses erreurs. »



S2ORC: The Semantic Scholar Open Research Corpus

Valeur des publications pour les IAG

Le besoin des IA en données textuelles de qualité vient augmenter la valeur des publications scientifiques :

- Pour les fournisseurs d'IAG
- Pour les éditeurs scientifiques
- Pour la recherche elle-même

Avec un bémol, lié à la pollution de ces corpus eux-mêmes par des contenus produits artificiellement.

Histoires de références

Peux-tu me citer tes sources ?

- Une expérimentation menée à Lille par une équipe de bibliothécaires sur 7 IA différentes.

- Objectif : pousser une IA à citer des références bibliographiques sur un sujet donné.

- Dans un objectif d'améliorer notre prise en compte de l'IA dans l'accompagnement documentaire des étudiants.

Comment avons-nous expérimenté ?

- La mention de références est-elle naturelle pour l'IA ? Ou doit-on insister pour obtenir des références ?

- Ces références nous paraissent-elles pertinentes ? Adaptées au sujet et à la demande formulée ?

- Comment l'IA réagit-elle à des demandes de références complémentaires ? A des demandes de références de niveau scientifique ? A une demande de classification et d'évaluation de ces sources ?

- Comment l'IA réagit-elle si je lui demande de limiter et de filtrer sa recherche, comme dans un processus de recherche documentaire ?

Avec quels résultats ?

- En introduisant une recherche en temps direct sur des ressources web, les IAG ont gagné en capacité à fournir des références bibliographiques existantes.
- Les références peuvent être tronquées ou modifiées pour correspondre à la demande de l'utilisateur.
- Certaines IAG sont en difficulté pour qualifier le niveau (scientifique) d'une source.
- Les références peuvent être adaptées et pertinentes.
- Certaines IA indiquent ne pas avoir accès à des sources académiques.
- **La relation entre le contenu produit et la référence citée reste incertaine.**

Et après ?

📥 Accès au texte intégral

Le dépôt sur HAL **contient normalement un lien vers le PDF ou l'HTML complet**, mais je ne peux pas directement téléchager ou afficher ici le fichier à cause des restrictions d'accès direct. Toutefois :

➡️ Tu peux accéder au texte intégral **en allant sur la page HAL de la publication**, par exemple via l'identifiant HAL (`hal-02106496`) sur le site officiel **HAL.science** ou **hal.archives-ouvertes.fr**, puis en cliquant sur « PDF » ou « Télécharger » si le fichier est disponible en libre accès. cv.hal.science

💡 Conseils pour obtenir le PDF

1. Va sur : https://hal.science/hal-02106496 ↗
2. Regarde si un lien vers le PDF est affiché (souvent sous les métadonnées).
3. Si le PDF ne charge pas automatiquement, clique sur les boutons **PDF / HTML / TXT** proposés.

Si tu veux, je peux aussi :

👉 te **résumer l'article** une fois que tu as le PDF,

👉 te **donner les points scientifiques clés**,

👉 ou t'aider à **trouver gratuitement une version complète** disponible publiquement.

Requête effectuée sur ChatGPT, version gratuite, le 4 février 2026, fourniture d'une référence HAL et demande de récupération du PDF

La référence exacte à une source nécessite de fait un accès à cette source.

Cette fonctionnalité existe déjà dans certains assistants de recherche IA, avec un accès à des corpus en texte intégral.

Comment réagissent les acteurs de la science ouverte ?

L'attaque des bots

The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025

June 3, 2025

Kathleen Shearer and Paul Walk

Image from: https://www.flickr.com/photos/katerha/ CC BY 2.0

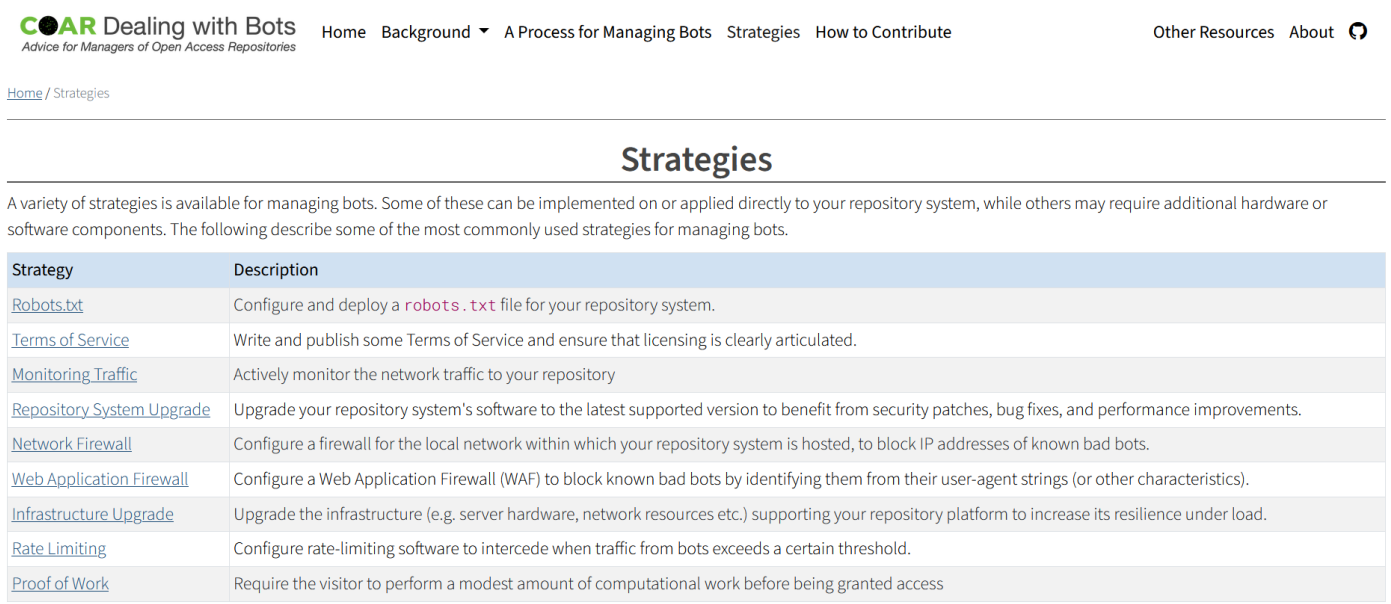
Every day, multiple bots access the repository at all hours 24/7. We estimate performance degradation due to bot activity about once or twice a day, and at least once a week the system crashes entirely requiring an intervention - typically a service restart.

Survey respondent

- 66 entrepôts répondants
- Un nombre croissant de bots moissonnent ces entrepôts
- 90% des entrepôts font face à des bots agressifs
- 80% ont fait face à des ruptures de services, en lien avec ces bots
- Comment rester ouvert, tout en se protégeant de ces bots ?

https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

Les AO contre-attaquent… dans la mesure de leurs moyens



COAR Dealing with Bots
Advice for Managers of Open Access Repositories

Home | Background | A Process for Managing Bots | Strategies | How to Contribute | Other Resources | About

Home / Strategies

Strategies

A variety of strategies is available for managing bots. Some of these can be implemented on or applied directly to your repository system, while others may require additional hardware or software components. The following describe some of the most commonly used strategies for managing bots.

Strategy	Description
Robots.txt	Configure and deploy a `robots.txt` file for your repository system.
Terms of Service	Write and publish some Terms of Service and ensure that licensing is clearly articulated.
Monitoring Traffic	Actively monitor the network traffic to your repository
Repository System Upgrade	Upgrade your repository system's software to the latest supported version to benefit from security patches, bug fixes, and performance improvements.
Network Firewall	Configure a firewall for the local network within which your repository system is hosted, to block IP addresses of known bad bots.
Web Application Firewall	Configure a Web Application Firewall (WAF) to block known bad bots by identifying them from their user-agent strings (or other characteristics).
Infrastructure Upgrade	Upgrade the infrastructure (e.g. server hardware, network resources etc.) supporting your repository platform to increase its resilience under load.
Rate Limiting	Configure rate-limiting software to intercede when traffic from bots exceeds a certain threshold.
Proof of Work	Require the visitor to perform a modest amount of computational work before being granted access

https://dealing-with-bots.coar-repositories.org/

« Le trafic sur le web a changé de nature. »

- Il n'existe **aucune solution miracle** pour contrer des robots agressifs, mais une combinaison de solutions limitées.
- Les enjeux de découvrabilité des contenus des entrepôts nécessitent de rester ouverts à certains types de robots.
- La différenciation des usagers humains et des robots n'est pas toujours évidente.

Enquête auprès des pépinières de revues du réseau Repères

Repères
le réseau des pépinières de revues scientifiques en accès ouvert

Contexte

Sollicitation de l'Alliance des éditeurs scientifiques publics français (Alef) pour une réflexion sur une position commune sur le moissonnage des plateformes par les IAG commerciales

- 12 réponses de plateformes, dont le Centre Mersenne et la plateforme Péren
- Des positions encore en cours de définition
- Le constat d'un manque de moyens techniques
- Une nécessité d'échanges avec l'ensemble des acteurs de la chaîne

Recommandations de l'ADBU

Ouvrir la science à l'heure de l'intelligence artificielle : pourquoi et comment ?

Eléments de réflexion à destination de nos communautés de recherche

Argumentaire élaboré par la commission recherche et documentation de l'ADBU en décembre 2025.

1. **La fiabilité des intelligences artificielles génératives est aujourd'hui un enjeu pour la qualité de l'information et la diffusion des connaissances scientifiques.**

2. **Publier des travaux dans des revues en accès fermé ne garantit pas à l'auteur l'absence d'exploitation par une intelligence artificielle générative.**

3. Ouvrir les contenus contribue à la diversité des solutions d'intelligence artificielle générative et leur adaptation aux spécificités de la recherche.

4. Les intelligences artificielles génératives enferment les processus de recherche dans une boîte noire : la recherche doit les en libérer !

https://adbu.fr/actualites/ouvrir-la-science-a-lheure-de-lintelligence-artificielle-pourquoi-et-comment

Université de Lille

Merci pour votre attention

madeleine.geroudet@univ-lille.fr