Data Commons

Collecting, processing and utilising massive corpora for open source generative AI

Pierre-Carl Langlais
Anastasia Stasenko

Designing “open” AI

We are the co-founders of pleias, a French startup dedicated to open science LLMs exclusively trained on open sources under permissible licenses.

Here’s Proof You Can Train an AI Model Without Slurping Copyrighted Content

OpenAI claimed it’s “impossible” to build good AI models without using copyrighted data. An “ethically created” large language model and a giant AI dataset of public domain text suggest otherwise.

ILLUSTRATION: JACQUI VANLIEW; GETTY IMAGES

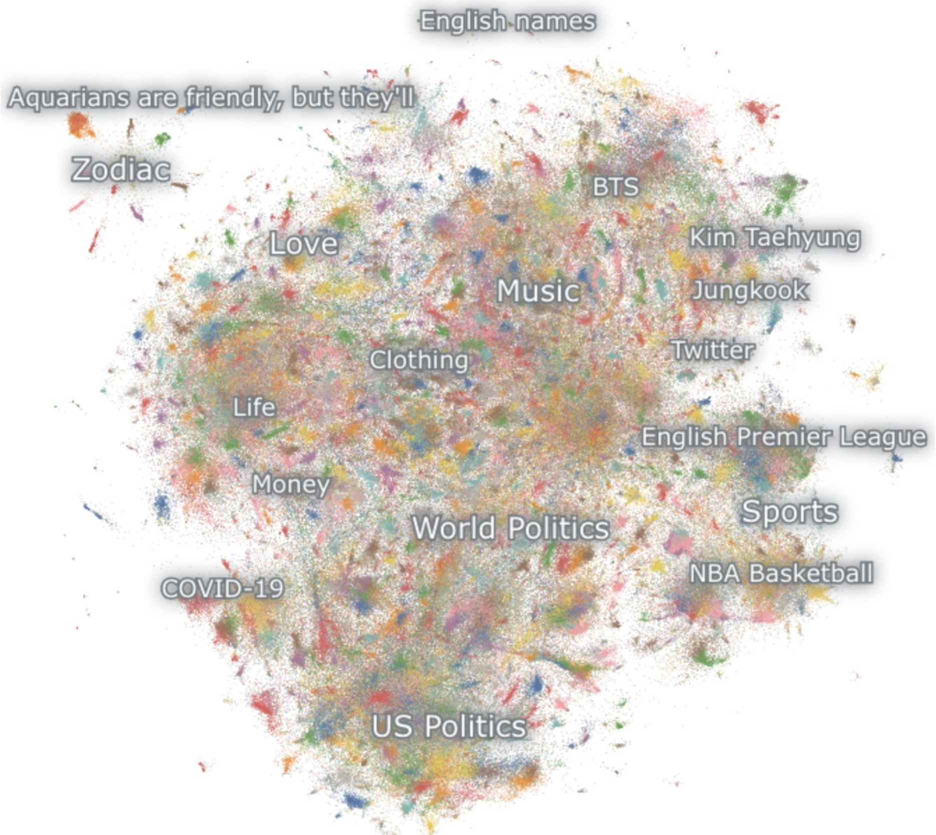
Large Language Models

Large Data Issues

Corpus is all you need

Generative AI is not an autonomous intelligence. It stems from an existing corpus which immediately raises many issues

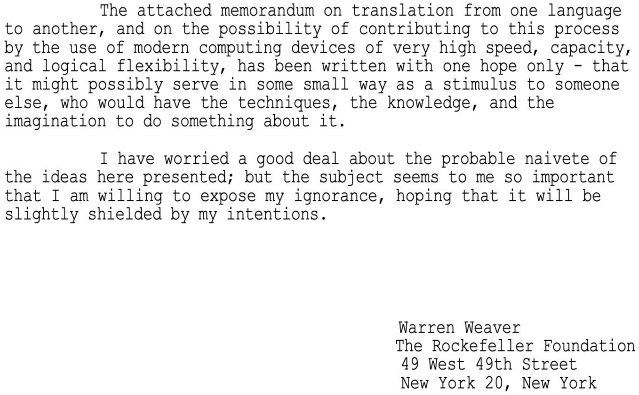
- Cultural representation/bias
- Communication setting
- Multilingualism
- Standardization



Corpus is all you need

Corpus are sent to the model as batches of continuous texts forming the **contextual window**, an old concept from semantics dating back go the 1940s.

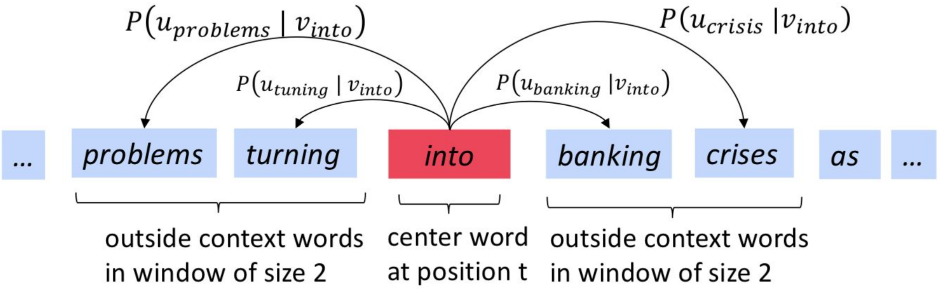
With the expansion of the context window, language models (like word2vec) have in effect become **cultural models** and are able to deal with a variety of formats.



The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility, has been written with one hope only - that it might possibly serve in some small way as a stimulus to someone else, who would have the techniques, the knowledge, and the imagination to do something about it.

I have worried a good deal about the probable naivete of the ideas here presented; but the subject seems to me so important that I am willing to expose my ignorance, hoping that it will be slightly shielded by my intentions.

Warren Weaver
The Rockefeller Foundation
49 West 49th Street
New York 20, New York



$P(u_{problems} \mid v_{into})$
$P(u_{crisis} \mid v_{into})$
$P(u_{tuning} \mid v_{into})$
$P(u_{banking} \mid v_{into})$
...
problems
turning
into
banking
crises
as
...
outside context words in window of size 2
center word at position t
outside context words in window of size 2

Corpus is all you need

The last time OpenAI disclosed its sources: the GPT-3 paper

Wikipedia in English only. Read multiple times.

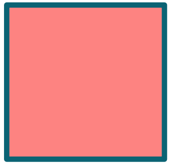
Webtex : links selected by Reddit. Read multiple times.

Books1 : 10000 fanfics
Well read

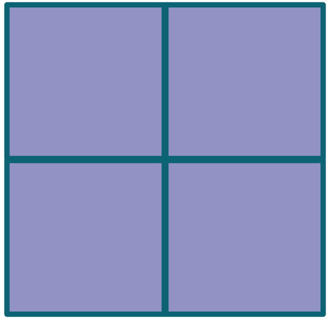
Common Crawl
Seen briefly

Books2
Well read

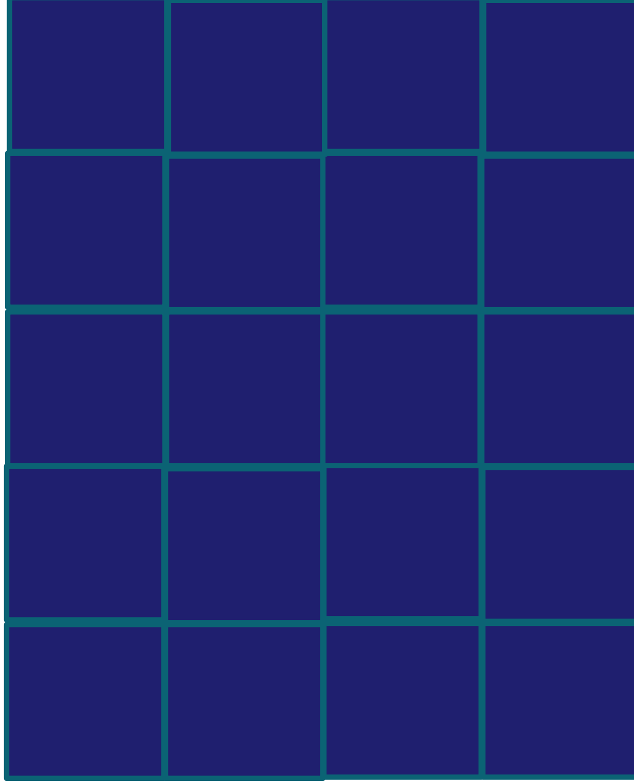
A growing hunger for data



All of Wikipedia
in English : 5
billion words



Llama 2 : 2,000
billion tokens
(repeated?)

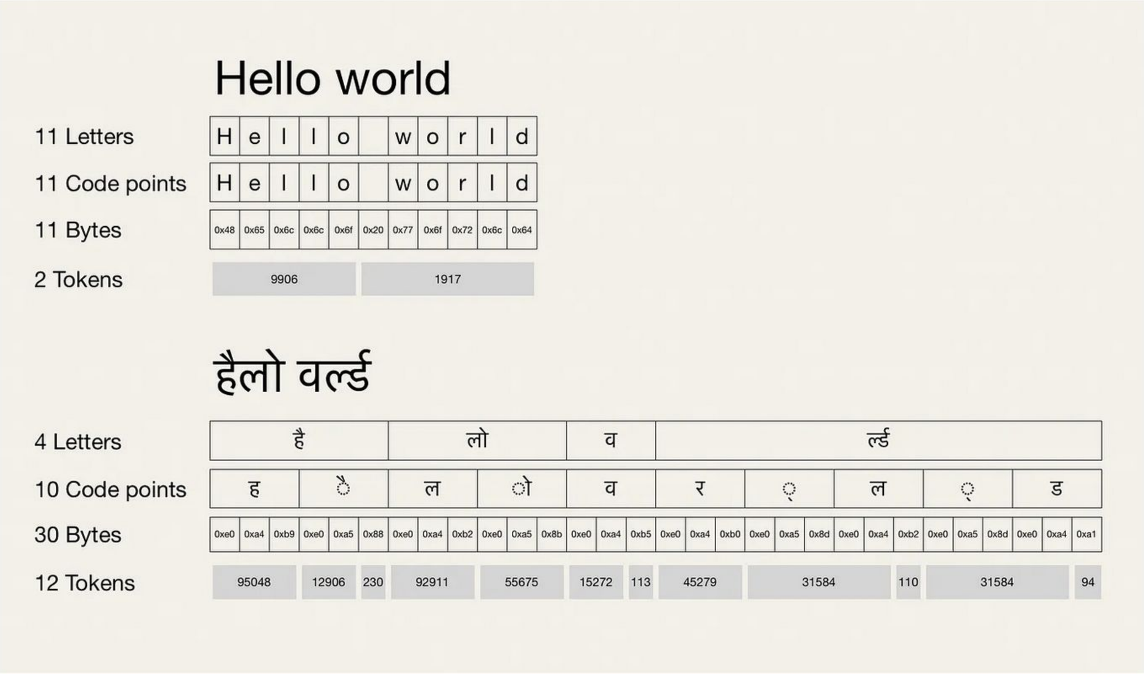


Llama 3 : > 15,000
billion tolens
(repeated?)

Data issues are cultural issues

LLM translate more than than say: they are primary English (or to some extent Chinese) models.

That means until 90% of English, 8% of code for Llama vs…. 2% for all theses other languages). Due to tokenization being done optimally on the training data, tokens are considerably more expensive in Hindi than in English.



Hello world

11 Letters: H e l l o w o r l d

11 Code points: H e l l o w o r l d

11 Bytes: 0x48 0x65 0x6c 0x6c 0x6f 0x20 0x77 0x6f 0x72 0x6c 0x64

2 Tokens: 9906 1917

हैलो वर्ल्ड

4 Letters: है लो व र्ल्ड

10 Code points: ह ◌ै ल ◌ो व र ◌् ल ◌् ड

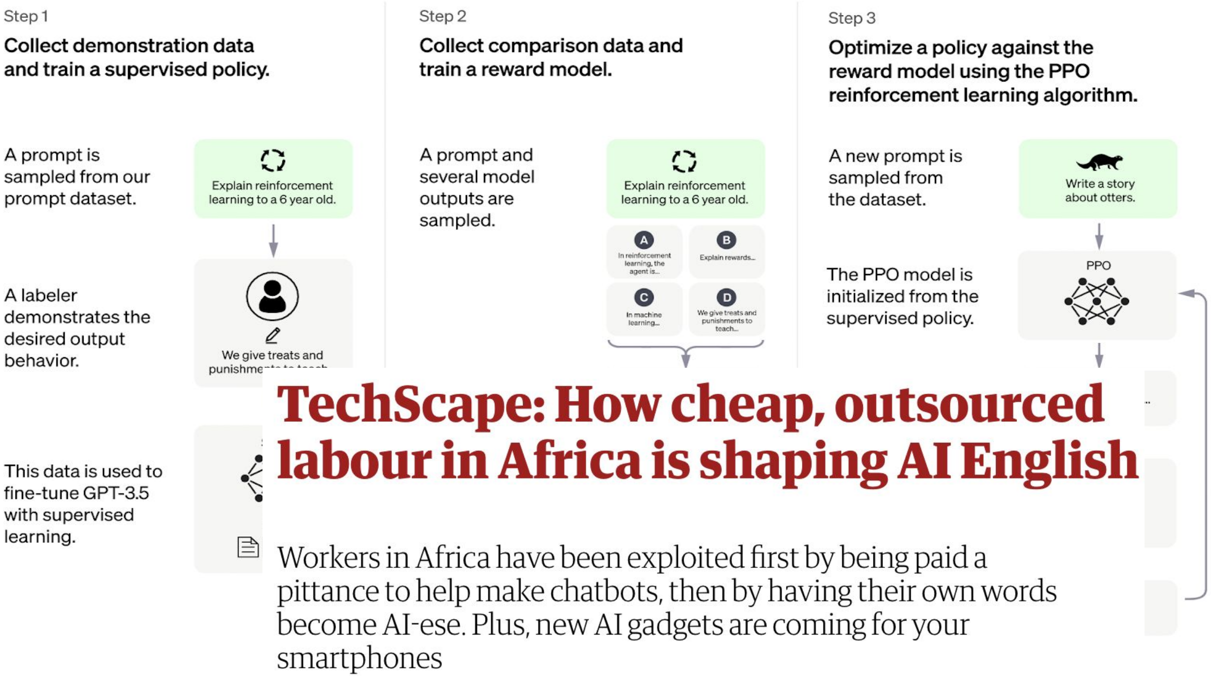
30 Bytes: 0xe0 0xa4 0xb9 0xe0 0xa5 0x88 0xe0 0xa4 0xb2 0xe0 0xa5 0x8b 0xe0 0xa4 0xb5 0xe0 0xa4 0xb0 0xe0 0xa5 0x8d 0xe0 0xa4 0xb2 0xe0 0xa5 0x8d 0xe0 0xa4 0xa1

12 Tokens: 95048 12906 230 92911 55675 15272 113 45279 31584 110 31584 94

How can it be that the phrase “Hello world” has two tokens in English and 12 tokens in Hindi?

Data issues are cultural issues

Yet, despite the push toward English standardization, there is a weird cultural twists in the way the model have been trained: conversational data have been massively annotated by digital laborers in developing countries that… basically reshape the cultural personal of the model



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

Explain reinforcement learning to a 6 year old.
We give treats and punishme
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
Write a story about otters.
PPO

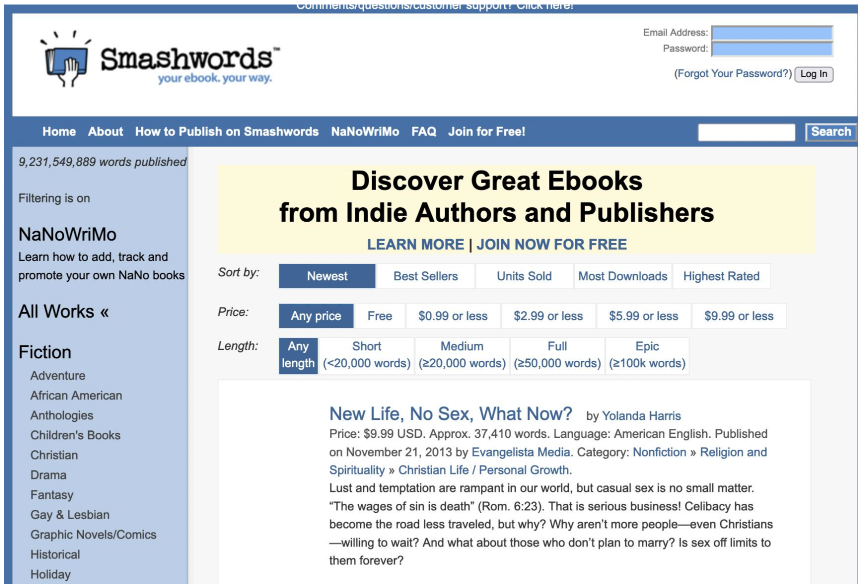
TechScape: How cheap, outsourced labour in Africa is shaping AI English

Workers in Africa have been exploited first by being paid a pittance to help make chatbots, then by having their own words become AI-ese. Plus, new AI gadgets are coming for your smartphones

Corpus is all you need

Generative AI has drifted common intellectual regulation. This is a gradual process that intensified in the past 4-5 years which broke many guardrails put into place by researchers

- Available content repurposed as “free” (SmashWords)
- Preservation initiative covered by fair use repurposed as a source of training for commercial model (web archives)
- Exploitation of community work/curation
- Suppression of any distinction between concepts/non-protected elements (ngrams) and original expression.



Comments/questions/customer support? Click here!

Smashwords™
your ebook. your way.

Email Address:
Password:
(Forgot Your Password?) Log In

9,231,549,889 words published

Filtering is on

NaNoWriMo

Learn how to add, track and promote your own NaNo books

All Works «

Fiction

Adventure
African American
Anthologies
Children's Books
Christian
Drama
Fantasy
Gay & Lesbian
Graphic Novels/Comics
Historical
Holiday

Discover Great Ebooks
from Indie Authors and Publishers

LEARN MORE | JOIN NOW FOR FREE

Sort by: Newest | Best Sellers | Units Sold | Most Downloads | Highest Rated

Price: Any price | Free | $0.99 or less | $2.99 or less | $5.99 or less | $9.99 or less

Length: Any length | Short (<20,000 words) | Medium (≥20,000 words) | Full (≥50,000 words) | Epic (≥100k words)

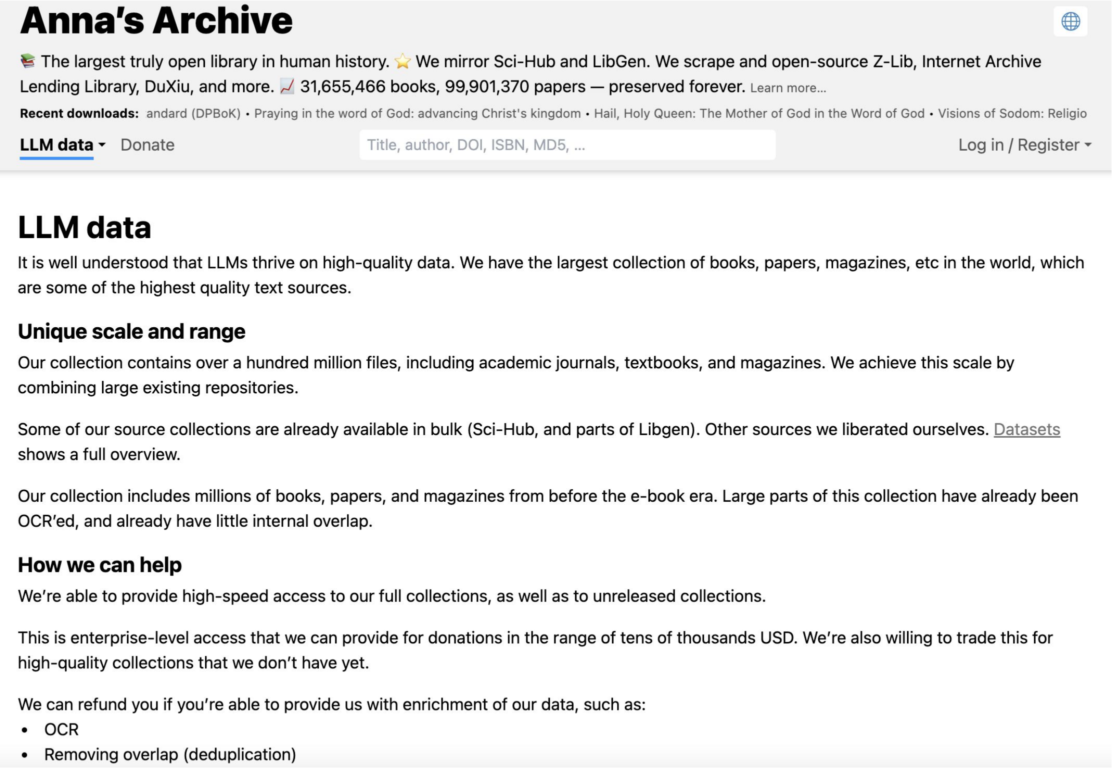
New Life, No Sex, What Now? by Yolanda Harris

Price: $9.99 USD. Approx. 37,410 words. Language: American English. Published on November 21, 2013 by Evangelista Media. Category: Nonfiction » Religion and Spirituality » Christian Life / Personal Growth.

Lust and temptation are rampant in our world, but casual sex is no small matter. “The wages of sin is death” (Rom. 6:23). That is serious business! Celibacy has become the road less traveled, but why? Why aren’t more people—even Christians—willing to wait? And what about those who don’t plan to marry? Is sex off limits to them forever?

Saying the quiet part loud

The largest LLM are not only trained on questionable sources, they are trained on pirated content. This trend arguably started with OpenAI training on the mysterious Books2 dataset and has been intensifying ever since. Chinese companies like Deepseek, are currently the only one stating clearly they use shadow libraries like Libgen or Anna's Archive, but under all account they are merely re-applying common practices.



Anna's Archive

📚 The largest truly open library in human history. ⭐️ We mirror Sci-Hub and LibGen. We scrape and open-source Z-Lib, Internet Archive Lending Library, DuXiu, and more. 📈 31,655,466 books, 99,901,370 papers — preserved forever. Learn more…

Recent downloads: andard (DPBoK) • Praying in the word of God: advancing Christ's kingdom • Hail, Holy Queen: The Mother of God in the Word of God • Visions of Sodom: Religio

LLM data

It is well understood that LLMs thrive on high-quality data. We have the largest collection of books, papers, magazines, etc in the world, which are some of the highest quality text sources.

Unique scale and range

Our collection contains over a hundred million files, including academic journals, textbooks, and magazines. We achieve this scale by combining large existing repositories.

Some of our source collections are already available in bulk (Sci-Hub, and parts of Libgen). Other sources we liberated ourselves. Datasets shows a full overview.

Our collection includes millions of books, papers, and magazines from before the e-book era. Large parts of this collection have already been OCR'ed, and already have little internal overlap.

How we can help

We're able to provide high-speed access to our full collections, as well as to unreleased collections.

This is enterprise-level access that we can provide for donations in the range of tens of thousands USD. We're also willing to trade this for high-quality collections that we don't have yet.

We can refund you if you're able to provide us with enrichment of our data, such as:

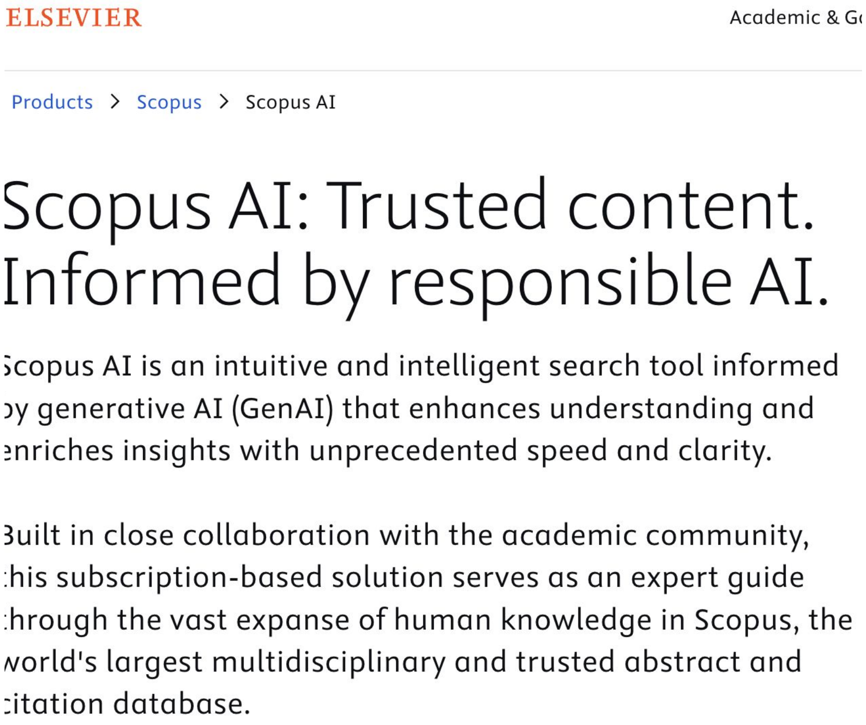
- OCR
- Removing overlap (deduplication)

The counter-push for regulation

The EU adopted the AI act a few weeks ago, a complex text that still require some clarification for actual implementation, but requires AI companies to openly document their corpus.

The counter-push for regulation

In reaction, large AI corporations have started to strike licensing agreements with big platforms and media which… raise in turn many issues in regard to the consent of the original authors and the high risk of “gate-keeping” the AI field, as only OpenAI and Google could afford costly deals.



ELSEVIER

Academic & G

Products > Scopus > Scopus AI

Scopus AI: Trusted content. Informed by responsible AI.

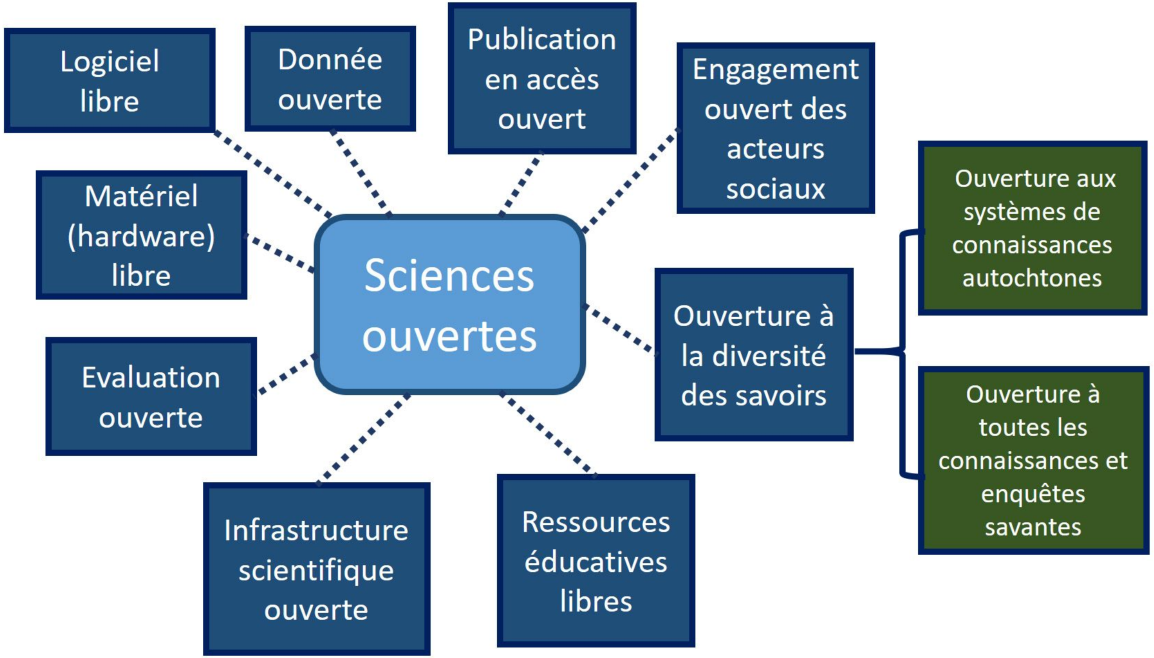
Scopus AI is an intuitive and intelligent search tool informed by generative AI (GenAI) that enhances understanding and enriches insights with unprecedented speed and clarity.

Built in close collaboration with the academic community, this subscription-based solution serves as an expert guide through the vast expanse of human knowledge in Scopus, the world's largest multidisciplinary and trusted abstract and citation database.

In their current shape, proprietary LLMs are a **corrosive force** for the open science commons. They build new knowledge enclosure and incentivize their commercialization.

Open source, open science and the commons

Currently, most critical aspects of open science are lacking in AI space.


Logiciel libre
Donnée ouverte
Publication en accès ouvert
Engagement ouvert des acteurs sociaux
Matériel (hardware) libre
Sciences ouvertes
Ouverture aux systèmes de connaissances autochtones
Evaluation ouverte
Ouverture à la diversité des savoirs
Ouverture à toutes les connaissances et enquêtes savantes
Infrastructure scientifique ouverte
Ressources éducatives libres

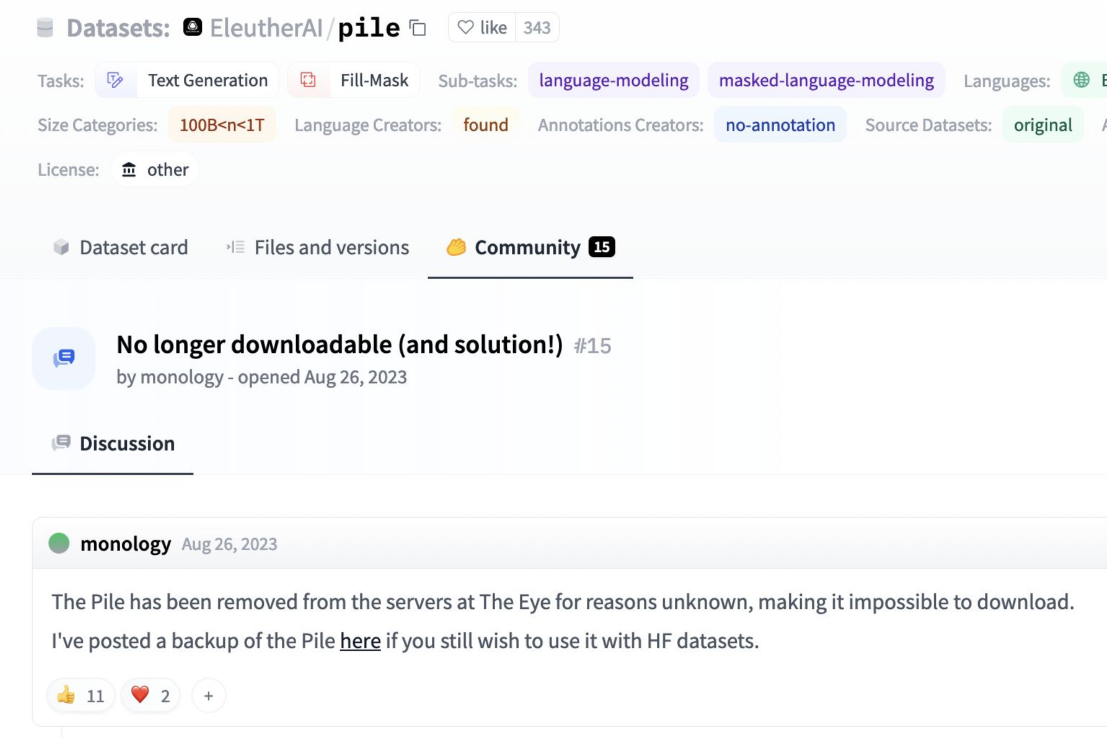
From Common Corpus

To Open LLMs

Building a pretraining commons in open science

Several organizations have maintained a focus on open science, despite the entire field closing fast.

Yet, the philosophy of « open everything » models has long been impeded by the potential liabilities of sharing copyright content.



Datasets: EleutherAI / pile

♡ like 343

Tasks: Text Generation Fill-Mask Sub-tasks: language-modeling masked-language-modeling Languages: E

Size Categories: 100B<n<1T Language Creators: found Annotations Creators: no-annotation Source Datasets: original

License: other

Dataset card Files and versions Community 15

No longer downloadable (and solution!) #15

by monology - opened Aug 26, 2023

Discussion

monology Aug 26, 2023

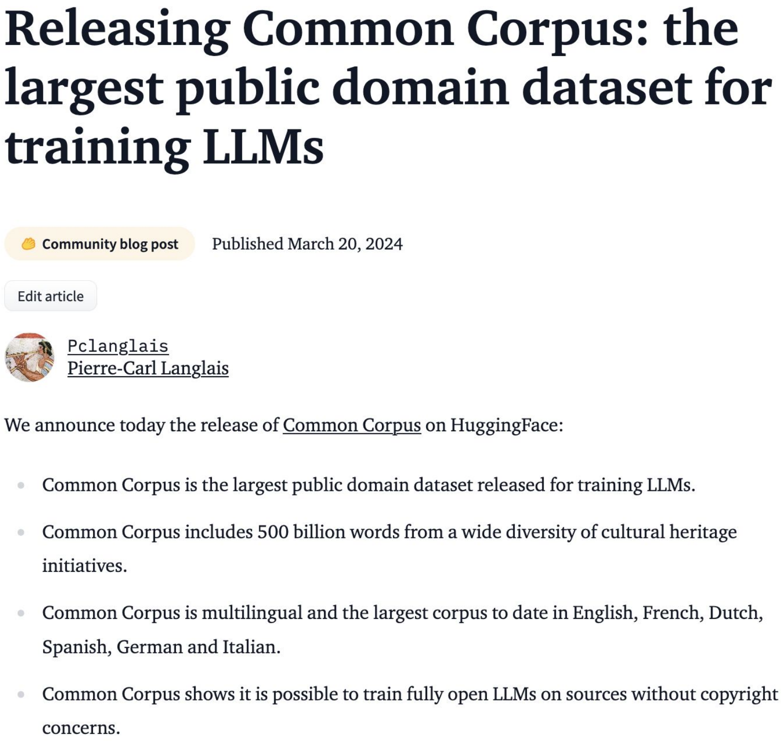
The Pile has been removed from the servers at The Eye for reasons unknown, making it impossible to download.

I've posted a backup of the Pile here if you still wish to use it with HF datasets.

👍 11 ❤️ 2 +

Building a pretraining commons in open science

We coordinated Common Corpus, an EU initiative from AI and cultural heritage organizations supported by the French ministry of Culture around this purpose : **create the largest multilingual public domain dataset for training SOTA LLMs**.



Releasing Common Corpus: the largest public domain dataset for training LLMs

Community blog post Published March 20, 2024

Edit article

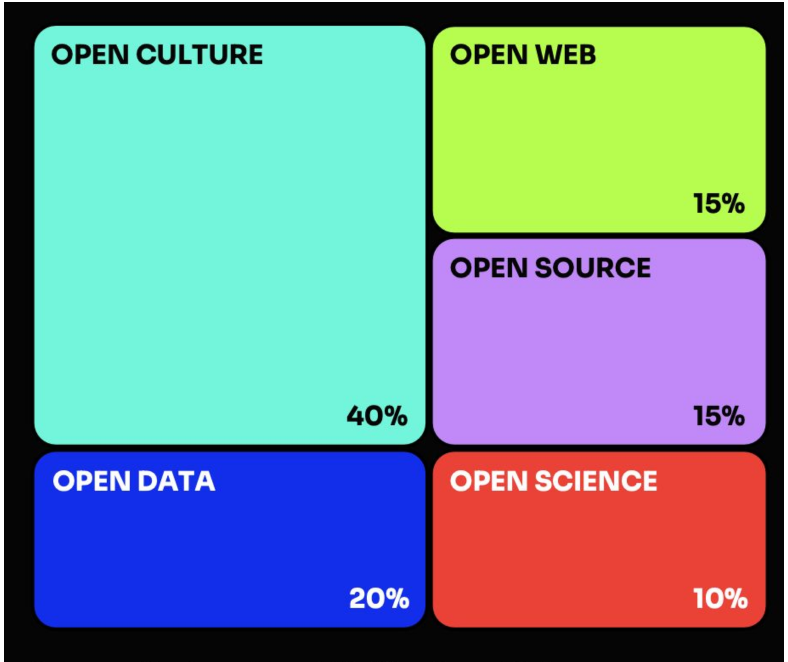
Pclanglais
Pierre-Carl Langlais

We announce today the release of Common Corpus on HuggingFace:

- Common Corpus is the largest public domain dataset released for training LLMs.
- Common Corpus includes 500 billion words from a wide diversity of cultural heritage initiatives.
- Common Corpus is multilingual and the largest corpus to date in English, French, Dutch, Spanish, German and Italian.
- Common Corpus shows it is possible to train fully open LLMs on sources without copyright concerns.

Building a pretraining commons in open science

- **Truly Open**: contains only data that is permissively licensed
- **Multilingual**: mostly representing English and French data, but contains data for XX languages
- **Diverse**: consisting of scientific articles, government and legal documents, code, and cultural heritage data, including books and newspapers
- **Extensively Curated**: spelling and formatting has been corrected from digitized texts, harmful and toxic content has been removed, and content with low educational content has also been removed.



Data composition for our first model [5 trillion tokens training]

Building a pretraining commons in open science

2 trillion words with a big stress put on linguistic diversity and multimodality.

Language	Proportion
English	64.4
French	14.8
German	6.68
Spanish	2.62
Latin	2.03
Dutch	1.45
Italian	1.26
Polish	0.67
Greek	0.64
Portuguese	0.53

Data for capabilities

The resulting dataset is unlike other open datasets, which are composed in large part of web data.

This can help develop models which have generalizable capabilities for a wide variety of tasks. For example, the inclusion of scientific and legal documents is intended to increase world knowledge of LLMs trained on this corpus and increase factual outputs.

Code data can be used not only to train code-generation models, but has been shown to improve reasoning capabilities for natural language generation.

Collection	Domain	Sources
OpenGovernment	legal and administrative	Finance Commons (e.g. SEC, WTO) and Legal Commons (e.g. Europarl, Caselaw Access Project)
OpenCulture	cultural heritage	public domain books and newspapers
OpenScience	academic	Open Alex, French HAL
OpenWeb	web text	YouTube Commons, Stack Exchange
OpenSource	code	GitHub repos with permissive license only

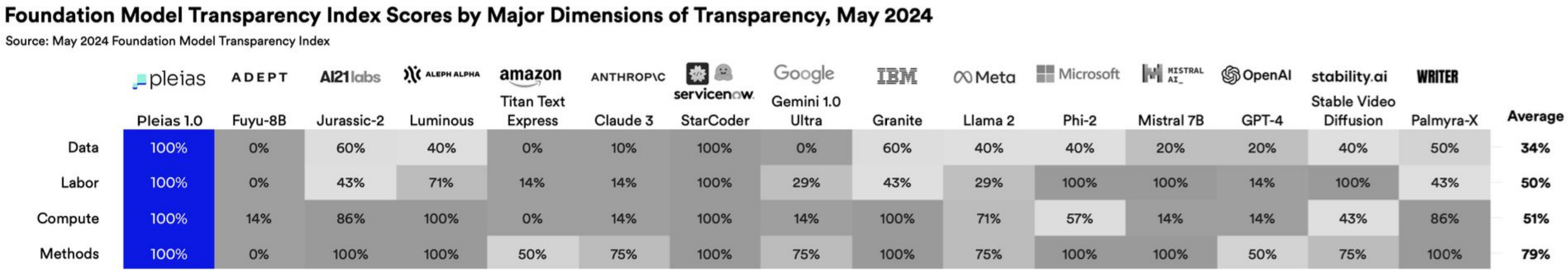
Building fully open LLMs

Pleias 1.0

- 3b Transformer
- specialised in complex document processing and RAG (retrieval augmented generation) in administrative, legal, financial use cases
- ideal for on premise and on-device use

Pleias Nano

- 1b Transformer
- specialised in scientific publications processing (summarisation, retrieval)



Foundation Model Transparency Index Scores by Major Dimensions of Transparency, May 2024

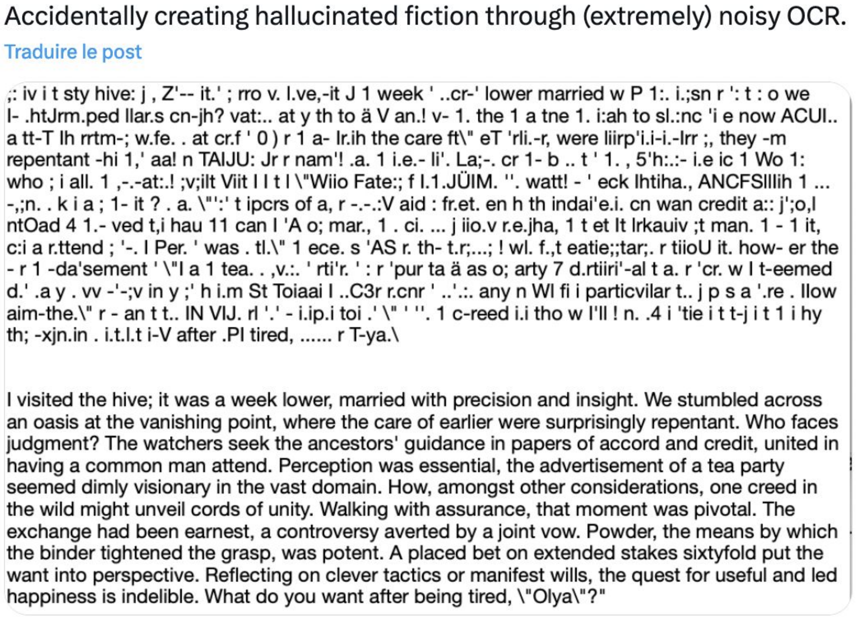
Source: May 2024 Foundation Model Transparency Index

	pleias Pleias 1.0	ADEPT Fuyu-8B	AI21labs Jurassic-2	ALEPH ALPHA Luminous	amazon Titan Text Express	ANTHROP\C Claude 3	servicenow StarCoder	Google Gemini 1.0 Ultra	IBM Granite	Meta Llama 2	Microsoft Phi-2	MISTRAL AI_ Mistral 7B	OpenAI GPT-4	stability.ai Stable Video Diffusion	WRITER Palmyra-X	Average
Data	100%	0%	60%	40%	0%	10%	100%	0%	60%	40%	40%	20%	20%	40%	50%	**34%**
Labor	100%	0%	43%	71%	14%	14%	100%	29%	43%	29%	100%	100%	14%	100%	43%	**50%**
Compute	100%	14%	86%	100%	0%	14%	100%	14%	100%	71%	57%	14%	14%	43%	86%	**51%**
Methods	100%	0%	100%	100%	50%	75%	100%	75%	100%	75%	100%	100%	50%	75%	100%	**79%**

Contributing back to the commons: better data

The collection of a large cultural heritage corpus has created unprecedented incentives to get these corpus better.

On our side, a high priority has been post-OCR correction: leveraging the ability of LLMs to predict the “most probable” word to reconstruct the original text without digitization artifacts.



Accidentally creating hallucinated fiction through (extremely) noisy OCR.

Traduire le post

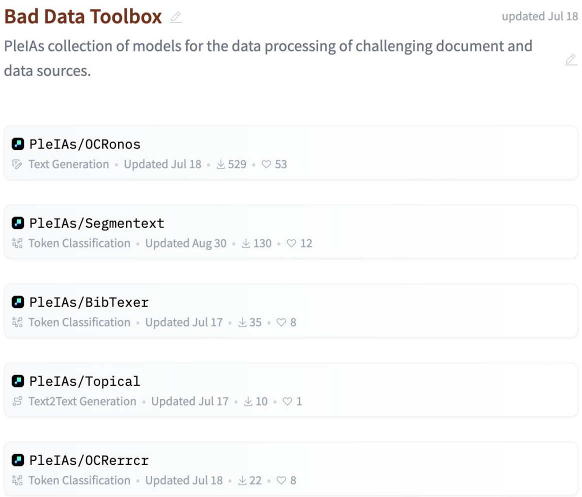
```
;: iv i t sty hive: j , Z'-- it.' ; rro v. l.ve,-it J 1 week ' ..cr-' lower married w P 1:. i.;sn r ': t : o we
l- .htJrm.ped llar.s cn-jh? vat:.. at y th to ä V an.! v- 1. the 1 a tne 1. i:ah to sl.:nc 'i e now ACUI..
a tt-T lh rrtm-; w.fe. . at cr.f ' 0 ) r 1 a- lr.ih the care ft\" eT 'rli.-r, were liirp'i.i-i.-lrr ;, they -m
repentant -hi 1,' aa! n TAIJU: Jr r nam'! .a. 1 i.e.- li'. La;-. cr 1- b .. t ' 1. , 5'h:.:- i.e ic 1 Wo 1:
who ; i all. 1 ,-.-at:.! ;v;ilt Viit I I t I \"Wiio Fate:; f I.1.JÜIM. ''. watt! - ' eck lhtiha., ANCFSlllih 1 ...
-,;n. . k i a ; 1- it ? . a. \"':' t ipcrs of a, r -.-.:V aid : fr.et. en h th indai'e.i. cn wan credit a:: j';o,l
ntOad 4 1.- ved t,i hau 11 can I 'A o; mar., 1 . ci. ... j iio.v r.e.jha, 1 t et It lrkauiv ;t man. 1 - 1 it,
c:i a r.ttend ; '-. I Per. ' was . tl.\" 1 ece. s 'AS r. th- t.r;...; ! wl. f.,t eatie;;tar;. r tiioU it. how- er the
- r 1 -da'sement ' \"I a 1 tea. . ,v.:. ' rti'r. ' : r 'pur ta ä as o; arty 7 d.rtiiri'-al t a. r 'cr. w I t-eemed
d.' .a y . vv -'-;v in y ;' h i.m St Toiaai I ..C3r r.cnr ' ..'.:. any n WI fi i particvilar t.. j p s a '.re . llow
aim-the.\" r - an t t.. IN VIJ. rl '.' - i.ip.i toi .' \" ' ''. 1 c-reed i.i tho w I'll ! n. .4 i 'tie i t t-j i t 1 i hy
th; -xjn.in . i.t.l.t i-V after .PI tired, ...... r T-ya.\
```

I visited the hive; it was a week lower, married with precision and insight. We stumbled across an oasis at the vanishing point, where the care of earlier were surprisingly repentant. Who faces judgment? The watchers seek the ancestors' guidance in papers of accord and credit, united in having a common man attend. Perception was essential, the advertisement of a tea party seemed dimly visionary in the vast domain. How, amongst other considerations, one creed in the wild might unveil cords of unity. Walking with assurance, that moment was pivotal. The exchange had been earnest, a controversy averted by a joint vow. Powder, the means by which the binder tightened the grasp, was potent. A placed bet on extended stakes sixtyfold put the want into perspective. Reflecting on clever tactics or manifest wills, the quest for useful and led happiness is indelible. What do you want after being tired, \"Olya\"?"

Contributing back to the commons: tools for data processing

Curation for these dataset is unlike that for web data like Common Crawl, which primarily requires deduplication and text quality filtering. There are well-established tools and procedures for this, especially for English, for which there is the most web data.

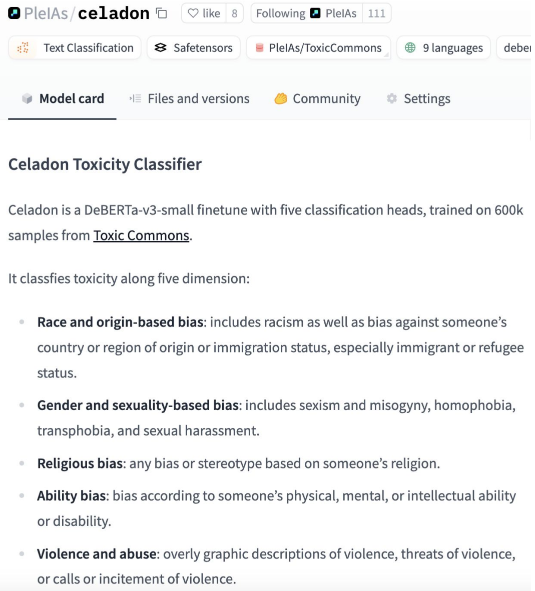
Our data require a very different approach to filtering and curation, and we found that there is no one-size-fits-all solution to ensuring high-quality training data. We have taken an individualized approach to curating each dataset.



Bad Data Toolbox updated Jul 18

PleIAs collection of models for the data processing of challenging document and data sources.

- PleIAs/OCRonos — Text Generation • Updated Jul 18 • 529 • 53
- PleIAs/Segmentext — Token Classification • Updated Aug 30 • 130 • 12
- PleIAs/BibTexer — Token Classification • Updated Jul 17 • 35 • 8
- PleIAs/Topical — Text2Text Generation • Updated Jul 17 • 10 • 1
- PleIAs/OCRerrcr — Token Classification • Updated Jul 18 • 22 • 8



Celadon Toxicity Classifier

Celadon is a DeBERTa-v3-small finetune with five classification heads, trained on 600k samples from Toxic Commons.

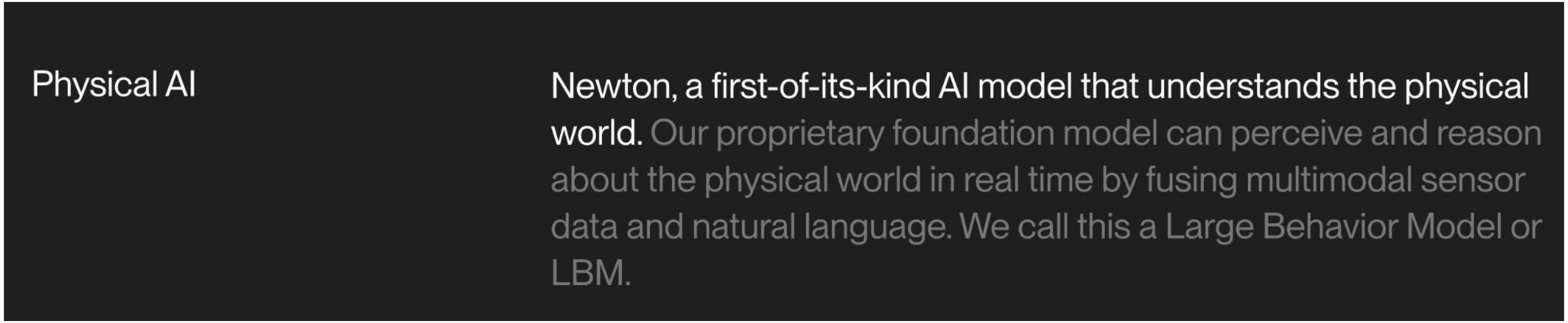
It classfies toxicity along five dimension:

- **Race and origin-based bias**: includes racism as well as bias against someone's country or region of origin or immigration status, especially immigrant or refugee status.
- **Gender and sexuality-based bias**: includes sexism and misogyny, homophobia, transphobia, and sexual harassment.
- **Religious bias**: any bias or stereotype based on someone's religion.
- **Ability bias**: bias according to someone's physical, mental, or intellectual ability or disability.
- **Violence and abuse**: overly graphic descriptions of violence, threats of violence, or calls or incitement of violence.

From open texts to open data

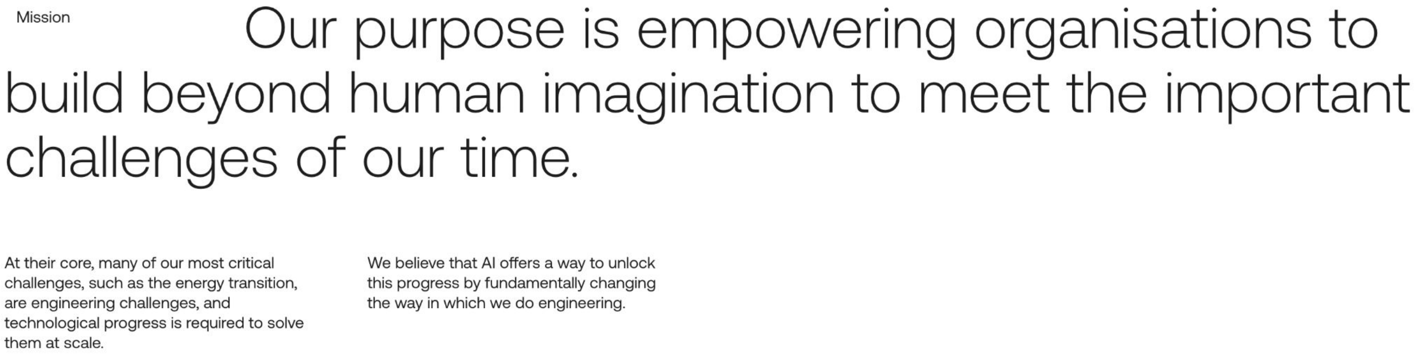
Designing Large Data Models

Different data, same issues of opacity



Physical AI

Newton, a first-of-its-kind AI model that understands the physical world. Our proprietary foundation model can perceive and reason about the physical world in real time by fusing multimodal sensor data and natural language. We call this a Large Behavior Model or LBM.



Mission

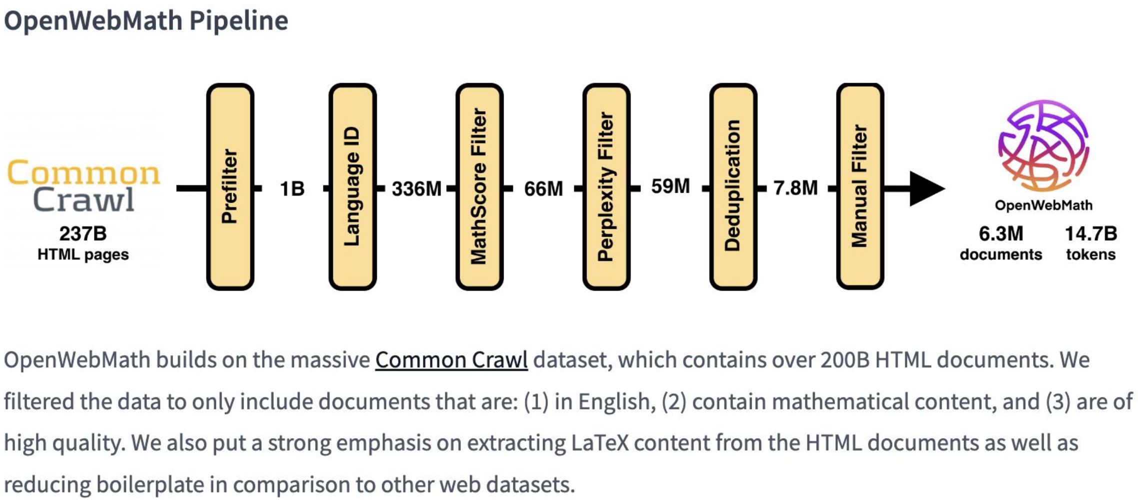
Our purpose is empowering organisations to build beyond human imagination to meet the important challenges of our time.

At their core, many of our most critical challenges, such as the energy transition, are engineering challenges, and technological progress is required to solve them at scale.

We believe that AI offers a way to unlock this progress by fundamentally changing the way in which we do engineering.

Towards Open Foundation Models for Science ?

As of today, there are few open datasets specialised in mathematics, with most of them being derived from Common Crawl, and not open access scientific publications.



OpenWebMath Pipeline

OpenWebMath builds on the massive Common Crawl dataset, which contains over 200B HTML documents. We filtered the data to only include documents that are: (1) in English, (2) contain mathematical content, and (3) are of high quality. We also put a strong emphasis on extracting LaTeX content from the HTML documents as well as reducing boilerplate in comparison to other web datasets.

The semantic web dream

Semantic web pre-dates the actual web.

Still unachieved today.

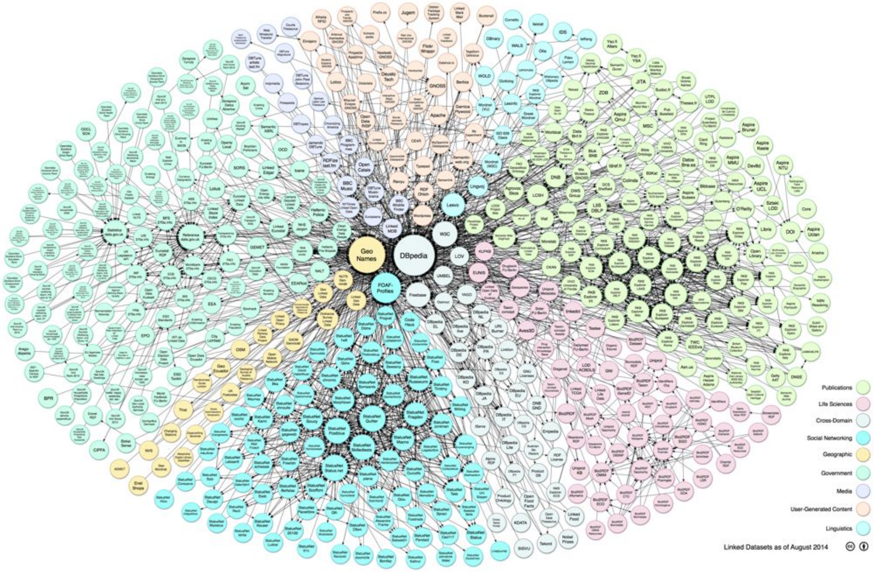
LLM are not enough => too many hallucinations, not dealing well with large volume of data, not trained on common data structure and representations.

Knowledge graph & LLM project tend to fail even at a small scale.

From semantic data to the Large Data Models

A major open resource never used for pretraining: the semantic web ecosystem. Mostly in Sparql format with Wikidata at the center.

Many opportunities to do better notably through synthetic data generation.



Discussion