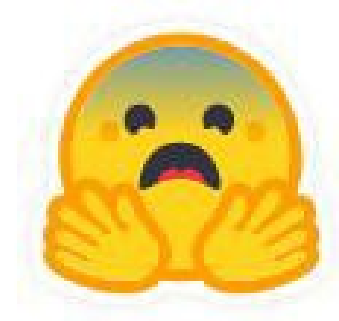


Danish Rights Alliance's Books3 case

Thomas Heldrup,
Head of Content Protection & Enforcement





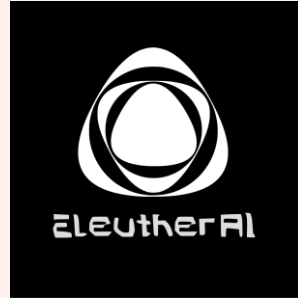
404

Sorry, we can't find the page you are looking for.

Data creators



Data collectors



Data hosts



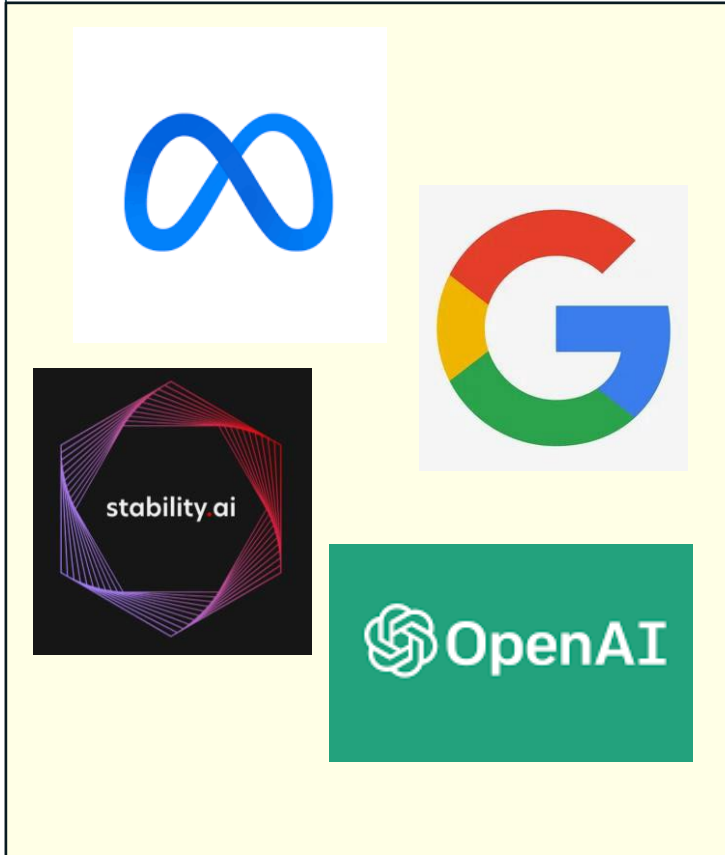
kaggle

Data creators: Creatives who produce works or creations.

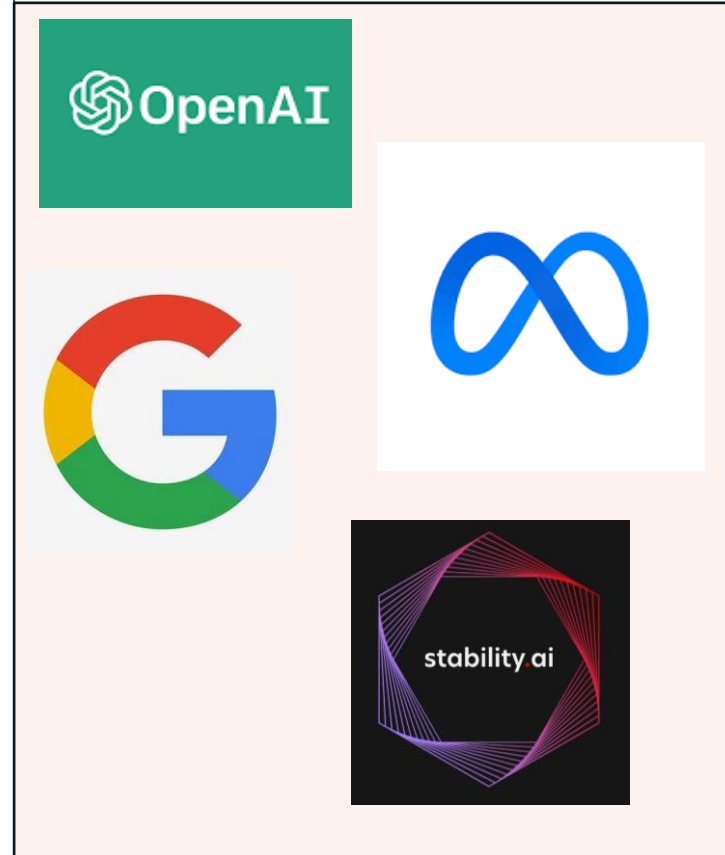
Data collectors: Individuals, organizations, or businesses that collect data (including works and/or creations) in a dataset, which can be used to train artificial intelligence.

Data hosts: Platforms that store and distribute datasets that can be used to train artificial intelligence.

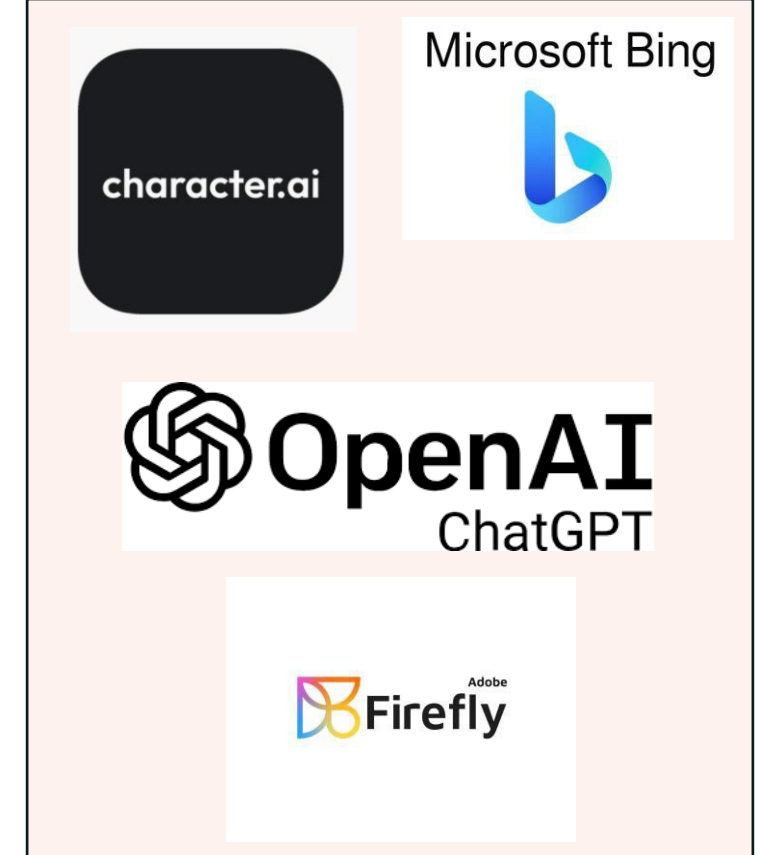
Foundation model creator



Foundation model provider



Generative AI service provider

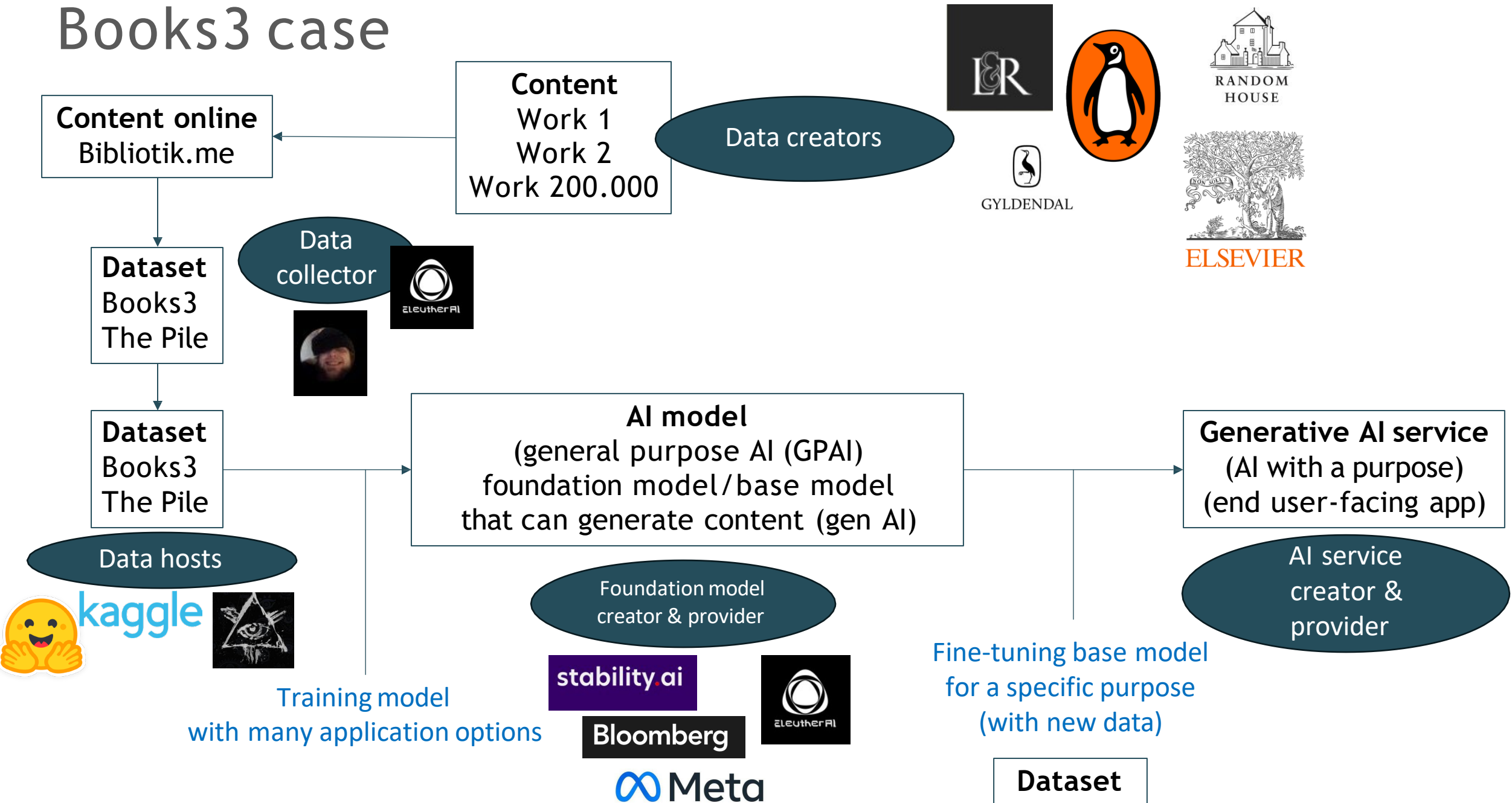


Foundation model creator: Organizations, businesses, individual(s) who develop a foundational model with multiple applications, including generating content such as images, text, and audio (generative AI).

Foundation model provider: Organizations, businesses, individual(s) who provide others with access to a base model (via API or open-source download).

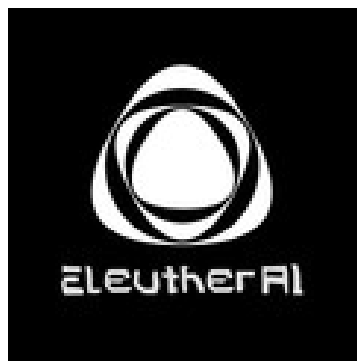
Generative AI service provider: Fine-tunes a foundation model with new data for a generative artificial intelligence service and offers it to others (via API or open-source download).

Books3 case



Books3 - Results and experiences

Distribution
(data
collectors
and hosts)



kaggle

AI model
creators
&
providers



stability.ai

Bloomberg



Books3 - Discovery

Thread

Shawn Presser @theshawwn

Suppose you wanted to train a world-class GPT model, just like OpenAI. How? You have no data.

Now you do. Now everyone does.

Presenting "books3", aka "all of bibliotik"

- 196,640 books
- in plain .txt
- reliable, direct download, for years: the-eye.eu/public/AI/pile...



Index of /public/AI/pile_preliminary_compon

File Name	Date Modified	Size	Kind
2020-09-08-arxiv-extracts-nofallback-until-2007...	01-Jan-1970 00:00	16G	
EuroParliamentProceedings_1996_2011.jsonl.zst	01-Jan-1970 00:00	1G	
FreeLaw_Opinions.jsonl.zst	01-Jan-1970 00:00	16G	
Literotica.jsonl.zst	01-Jan-1970 00:00	4G	
NIH_EXPORTER_awarded_grant_text.jsonl.zst	01-Jan-1970 00:00	602M	
PMC_extracts.tar.gz	01-Jan-1970 00:00	26G	
PUBMED_title_abstracts_2019_baseline.jsonl.zst	01-Jan-1970 00:00	6G	
PhilArchive.jsonl.zst	01-Jan-1970 00:00	761M	
books1.tar.gz	01-Jan-1970 00:00	2G	
books3.tar.gz	01-Jan-1970 00:00	37G	
github.tar	01-Jan-1970 00:00	106G	
hn.tar.gz	01-Jan-1970 00:00	674M	
openwebtext2.jsonl.zst.tar	01-Jan-1970 00:00	27G	
pile_uspto.tar	01-Jan-1970 00:00	11G	
stackexchange_dataset.tar	01-Jan-1970 00:00	34G	
ubuntu IRC until 2020_9_1.jsonl.zst	01-Jan-1970 00:00	2G	



```
Opium - Elsebeth Egholm.epub.txt
Opium
_Af samme forfatter_
De Frie Kvinders Klub, 1999
Scirocco, 2000
Mig og min ø, 2000
Skjulte fejl og mangler, 2002
Elsebeth Egholm
Opium
Lindhardt og Ringhof
_Opium_
© Elsebeth Egholm 2001
Lindhardt og Ringhof Forlag
Forside: Imperiet
2. udgave, 2. oplag
ISBN 978-87-11-41427-9
www.lrforglag.dk
_Til Fruake, Kirsten og Lise-Lotte
- og for de næste tyve år...
"If you wake at midnight, and hear a horse's feet,
Don't go drawing back the blind, or looking in the
street,
Them that asks no questions isn't told a lie.
Watch the wall, my darling, while the Gentlemen go
```



```
from datasets import load_dataset.py x
Users > thomash > from datasets import load_dataset.py > ...
1 from datasets import load_dataset
2 books_dataset = load_dataset('the_pile_books3')
3 filtered = books_dataset['train'].filter(lambda row: row['text'].find('L
4 print(filtered)
5 for row in filtered:
6     print([row['title']])

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
Found cached dataset the_pile_books3 (/Users/thomash/.cache/huggingface/datasets/
100%|
Dataset({
  features: ['title', 'text'],
  num_rows: 64
})
Opium - Elsebeth Egholm
Operation Firefight - Chris Ryan
Ebbe Larsen-Dengang i Danmark
En anarkists minnen - Petr Alekseevic Kropotkin
Hafnia punk - Benn Q
Hævnens time - Chris Ryan
Hellberg, Hans-Eric - Kram
Var tids hjalte - Michail Lermontov
Veninder - Fay Weldon
Førch - Michael Teschl
```



books3

Name	Date Modified	Size	Kind
A Kite in the evening sky - Snaik Kaqir.epub.txt	4 September 2020 at 11:49	239 KB	Plain Text
A Knife in the Fog - Bradley Harper (retail).epub.txt	4 September 2020 at 08:20	482 KB	Plain Text
A Knight in Tarnished Armor - Jill Barnett.epub.txt	4 September 2020 at 09:50	187 KB	Plain Text
A Knight of the Word - Terry Brooks.epub.txt	4 September 2020 at 09:49	594 KB	Plain Text
A Knight to Remember - Christina Dodd.epub.txt	4 September 2020 at 09:20	540 KB	Plain Text
A Knock on the Door - Phil FontaineAimee Craft.epub.txt	4 September 2020 at 11:48	624 KB	Plain Text
A L Bird - The Good Mother (retail) (epub).epub.txt	4 September 2020 at 08:40	379 KB	Plain Text
A La Mod - My So-Called Tranquil Family Life in Rural France.epub.txt	4 September 2020 at 08:04	491 KB	Plain Text
A la Mode - Mark Scarbrough.epub.txt	4 September 2020 at 13:24	423 KB	Plain Text
A la poursuite de Seth - J.R. Loveless.epub.txt	4 September 2020 at 07:50	499 KB	Plain Text
A la recherche d'une musique co - Pierre Schaeffer.epub.txt	4 September 2020 at 10:13	432 KB	Plain Text
A la recherche du reel - d'Espagnat.epub.txt	4 September 2020 at 10:34	585 KB	Plain Text
A la recherche du temps perdu 6 - Marcel PROUST.epub.txt	4 September 2020 at 10:42	943 KB	Plain Text
A la sombra de los cuervos - Javier Rebolledo.epub.txt	4 September 2020 at 11:34	794 KB	Plain Text
A la vie - Unknown.epub.txt	4 September 2020 at 11:30	296 KB	Plain Text
A Lad of Grit_nodrm.epub.txt	4 September 2020 at 10:57	306 KB	Plain Text
A Ladder to the Sky - John Boyne.epub.txt	4 September 2020 at 09:01	635 KB	Plain Text
A Lady Awakened (Blackshear Family 1) - Cecilia Grant.epub.txt	4 September 2020 at 10:10	607 KB	Plain Text
A Lady Cyclist's Guide to Kashgar - Suzanne Joinson.epub.txt	4 September 2020 at 13:00	497 KB	Plain Text
A Lady in Shadown - Lene Kasberbol.epub.txt	4 September 2020 at 08:56	504 KB	Plain Text
A Lady Never Lies - Juliana Gray.epub.txt	4 September 2020 at 09:23	598 KB	Plain Text
A Lady Never Surrenders - Sabrina Jeffries.epub.txt	4 September 2020 at 10:33	539 KB	Plain Text
A Lady of Persuasion - Tessa Dare.epub.txt	4 September 2020 at 07:59	565 KB	Plain Text
A Lady Seduces - Bronwyn Scott.epub.txt	4 September 2020 at 07:36	118 KB	Plain Text
A Lady's Code of Misconduct - Meredith Duran.epub.txt	4 September 2020 at 13:33	583 KB	Plain Text
A Lady's Guide to Improper Behavior - Suzanne Enoch.epub.txt	4 September 2020 at 08:58	520 KB	Plain Text
A Lady's Guide to Skirting Scandal - Kelly Bowen.epub.txt	4 September 2020 at 13:07	156 KB	Plain Text
A Lady's Lesson in Scandal - Meredith Duran.epub.txt	4 September 2020 at 08:45	630 KB	Plain Text
A Lady's Wish - Katharine Ashe.epub.txt	4 September 2020 at 12:12	212 KB	Plain Text
A Ladybird First Grown-Up Picture Book (Ladybirds for Grown-Ups) - Jason Hazeley.epub.txt	4 September 2020 at 08:41	8 KB	Plain Text
A LaLa Land Addiction - Ashley Antoinette.epub.txt	4 September 2020 at 09:12	429 KB	Plain Text
A Lamp to Illuminate the Five Senses - Le Teonkhana.epub.txt	4 September 2020 at 11:21	1.4 MB	Plain Text

Books3 - Discovery

Index of /public/AI/pile/train/

..			
00.jsonl.zst	01-Jan-1970	00:00	14G
01.jsonl.zst	01-Jan-1970	00:00	14G
02.jsonl.zst	01-Jan-1970	00:00	14G
03.jsonl.zst	01-Jan-1970	00:00	14G
04.jsonl.zst	01-Jan-1970	00:00	14G
05.jsonl.zst	01-Jan-1970	00:00	14G
06.jsonl.zst	01-Jan-1970	00:00	14G
07.jsonl.zst	01-Jan-1970	00:00	14G
08.jsonl.zst	01-Jan-1970	00:00	14G
09.jsonl.zst	01-Jan-1970	00:00	14G
10.jsonl.zst	01-Jan-1970	00:00	14G
11.jsonl.zst	01-Jan-1970	00:00	14G
12.jsonl.zst	01-Jan-1970	00:00	14G
13.jsonl.zst	01-Jan-1970	00:00	14G
14.jsonl.zst	01-Jan-1970	00:00	14G
15.jsonl.zst	01-Jan-1970	00:00	14G
16.jsonl.zst	01-Jan-1970	00:00	14G
17.jsonl.zst	01-Jan-1970	00:00	14G
18.jsonl.zst	01-Jan-1970	00:00	14G
19.jsonl.zst	01-Jan-1970	00:00	14G
20.jsonl.zst	01-Jan-1970	00:00	14G
21.jsonl.zst	01-Jan-1970	00:00	14G
22.jsonl.zst	01-Jan-1970	00:00	14G
23.jsonl.zst	01-Jan-1970	00:00	14G
24.jsonl.zst	01-Jan-1970	00:00	14G
25.jsonl.zst	01-Jan-1970	00:00	14G
26.jsonl.zst	01-Jan-1970	00:00	14G
27.jsonl.zst	01-Jan-1970	00:00	14G
28.jsonl.zst	01-Jan-1970	00:00	14G
29.jsonl.zst	01-Jan-1970	00:00	14G



```
The Pile — unzstd 02.jsonl.zst — 80x24
Last login: Mon Aug 14 15:28:37 on ttys001

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
[MacBook-Air-2:The Pile thomash$ unzstd 02.jsonl.zst
02.jsonl.zst          : 4.60 GiB...
```



```
02.json
Hex Text Find Lindhardt Replace
Replace All Replace Replace & Find Previous Next
2354322300 om the copyright holder.", "meta": {"pile_set_name": "Pile-CC"}} {"text": " \nPetr Alekseevi\u010d K
2354322400 ropotkin\n\n# En anarkists minnen\n\nSaga Egmont\n\nEn anarkists minnen \n\u00e4r \u00f6versatt fr\u00
2354322500 0e5n ryska av Hellen Lindgren efter\n\nMemoirs of a Revolutionist\n\nCopyright \u00a9 1917, 2016 Pet
2354322600 r Alekseevi\u010d Kropotkin och SAGA Egmont, an imprint of Lindhardt og Ringhof A/S Copenhagen\n\nAl
2354322700 l rights reserved\n\nISBN: 9788711664001\n\n* * *\n\n1. e-bok version, 2016\n\nFormat: EPUB 3.0\n\n*
2354322800 * *\n\nSAGA Egmont www.saga-books.com \u0013 a part of Egmont, www.egmont.com\n\n# F\u00f6reta
2354322900 nStora anders sj\u00e4lvbiografi har i forna tider vanligen haft \n\u00e5gon av f\u00f6ljande tre typ
2354323000 er: s\u00e4r vilsefarande var jag; p\u00e4r det h\u00e4r s\u00e4ttet blev jag omv\u00e4nd. (S
2354323100 :t Augustinus.) S\u00e4r h\u00e4r d\u00e4r lig var jag; men vem \u00e4r kalla sig b\u00e4ttre! (Rous
2354323200 seau.) P\u00e4r detta s\u00e4tt danades \u00e5ngsamt inifr\u00e5n och av omst\u00e4ndigheternas g\u00e4
2354323300 t och n\u00e4r ett geni. (Goethe.)\n\nVid alla dessa former av sj\u00e4lvframst\u00e4llning \u00e4r
2354323400 f\u00f6rfattaren i \u00e4r sentlig grad sysselsatt med sig sj\u00e4lv.\n\nUnder det f\u00f6rflutna \u00
2354323500 00e5rhundradet pl\u00e4gade framst\u00e5ende personligheters sj\u00e4lvbiografi vara avfattad efter
2354323600 ettdera av f\u00f6ljande tvenne m\u00f6nster: S\u00e4r talangfull, s\u00e4r intagande var jag, s\u00e4r
2354323700 erk\u00e4nd och beundrad blev jag. (Johanne Luise Heiberg.) Eller: S\u00e4r talangfull och \u00e4rsk
2354323800 \u00e4rd var jag; s\u00e4r missk\u00e4nd blev jag och s\u00e4r h\u00e4r strider fick jag genom\u00
2354323900 e4mpa, innan jag vann ber\u00f6mmelsens krona. (H. C. Andersen.)\n\nI b\u00e5da dessa arter av levna
2354324000 dsbeskrivning har f\u00f6rfattaren f\u00f6ret\u00e4rdesvis sysselsatt sig med, vad hans medm\u00e4nn
2354324100 iskor hava tyckt och sagt om honom.\n\nF\u00f6rfattaren av f\u00f6religgande sj\u00e4lvbiografi har
2354324200 icke \u00e4gnat sin uppm\u00e4rksamhet \u00e4r s\u00e4r egna egenskaper, har icke heller skildrat n\u00e4r
2354324300 5gon sin kamp f\u00f6r att f\u00e4r dem erk\u00e4nda. \u00c4nnu mindre syssels\u00e4tter han sig med
2354324400 \u00e4rldens domslut; vad andra ha tyckt om honom, omn\u00e4mner han ej ens med ett ord.\n\nDet fin
2354324500 nes h\u00e4r ingen sj\u00e4lvbespeglning. F\u00f6rfattaren h\u00e4r icke till dem, som g\u00e4rna tal
2354324600 a om sig sj\u00e4lv; han g\u00e4r det motvilligt, med en viss blyghet. H\u00e4r finnes icke heller
2354324700 n\u00e4r s\u00e4r \u00e4rldens domslut; vad andra ha tyckt om honom, omn\u00e4mner han ej ens med ett ord.\n\nDet fin
2354324800 nes h\u00e4r ingen sj\u00e4lvbespeglning. F\u00f6rfattaren h\u00e4r icke till dem, som g\u00e4rna tal
```


Books3 - takedown procedure



From: DMCA/Abuse Agent
To: Thomas Heldrup
Attachments: image001.png, image002.png, image003.png

Hello Thomas,

Thank you for your reply on this issue.

I have been informed by the volunteer archivists that they were unaware of the fact that such works had been made publicly available and they immediately removed the offending data sets. This goes beyond the request you have made to some other files related to that project I have been told. This appears to be done in an abundance of caution rather than in response to specific requests or issues discovered with the additional data.

If you can please confirm that the offending links are not longer delivering data other than error codes such as 404 not found or 403 unauthorized and let me know that you are satisfied with the remedy, I would greatly appreciate it. Please do keep in mind that public search engine results may remain but are out of our control.

If you have any further questions about this case, please let me know by replying to this email.

Thanks,
James | The-Eye Volunteer Staff
Hardware and Platform Team

the-eye.eu/public/Al/pile_preliminary_components/books3.tar.gz

404 Not Found

nginx

Books3 - takedown procedure



Hugging Face Search models, datasets, users... Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

Datasets: the_pile_books3 like 61

Tasks: Text Generation Fill-Mask Sub-tasks: language-modeling masked-language-modeling Languages: English Multilinguality: monolingual Size Categories: 100K<n<1M Language Creators: found

Annotations Creators: no-annotation Source Datasets: original ArXiv: arxiv:2101.00027 License: mit

Dataset card Files and versions Community 4

The Dataset Viewer has been disabled on this dataset.

Dataset Card for the_pile_books3

Dataset Summary

This dataset is Shawn Presser's work and is part of EleutherAI/The Pile dataset.

This dataset contains all of bibliotik in plain .txt form, aka 197,000 books processed in exactly the same way as did for bookcorpusopen (a.k.a. books1). seems to be similar to OpenAI's mysterious "books2" dataset referenced in their papers. Unfortunately OpenAI will not give details, so we know very little about any differences. People suspect it's "all of libgen", but it's purely conjecture.

|download_size|36.8 Gib| |dataset_size|100.9 Gib|

Supported Tasks and Leaderboards

This dataset is used for Language Modeling.

Languages

The dataset is in English.

Dataset Structure

Data Instances

```
{'title': '07 LEGO Ninjago - The Search For Zane (Scholastic) - Kate Howard (retail)'}
'text': '\n\nTITLE PAGE\n\nFROM THE JOURNAL OF SENSEI GARMADON\n\nCHAPTER 1\n\nCHAPTER 2\n\nCHAPTER 3\n\n'
```

Data Fields

- title: title of the book
- text: text content of the book

Downloads last month **3,766**

Use in dataset library Edit dataset card

Evaluate models HF Leaderboard

Homepage: GitHub Paper: arXiv

Models trained or fine-tuned on the_pile_books3



- mosaicml/mpt-7b-storywriter**
Text Generation · Updated 7 da... · 19.3k · 635
- OccamRazor/mpt-7b-storywriter-4bit-12...**
Text Generation · Updated May 8 · 1.06k · 113
- TheBloke/MPT-7B-Storywriter-GGML**
Updated 27 days ago · 416 · 33
- jprafael/mpt-7b-instruct-sharded**
Text Generation · Updated May 23 · 277 · 2
- emozilla/mpt-7b-storywriter-fast**
Text Generation · Updated 27 days... · 272 · 11
- 4bit/mpt-7b-storywriter-4bit-128g**
Text Generation · Updated May 7 · 213 · 16

[Browse 11 models trained on this dataset](#)

Books3 - takedown procedure





Hugging Face

 **matthieumeeus** 14 days ago 

When attempting to download this with `datasets.load_dataset("the_pile_books3")`, I get the following error:

`FileNotFoundError: Couldn't find file at https://the-eye.eu/public/AI/pile_preliminary_components/books3.tar.gz`

Has the link been removed? How could I still download the dataset? Thanks :)


 **albertvillanova** 14 days ago 



Hi, [@matthieumeeus](#).




Mon, 24 Jul 2023 14:30:49 GMT




Indeed the data is no longer accessible: the host of the dataset has removed the data source files because of copyright issues with the authors of the books.



I am going to add a warning to the dataset card explaining this issue.


 **Hugging Face**



Datasets:  **monology/pile**  like 8

Tasks:  Text Generation Languages:  English License:  other

 Dataset card  Files  Community **3**

 **Report: Legal issue(s) #3**
by [julien-c](#)  - opened about 1 hour ago

 Discussion

 [julien-c](#)  about 1 hour ago · edited about 1 hour ago

HF received a notice of copyright infringement about this dataset from the Danish Rights Alliance due to its inclusion of copyrighted works of art inside the books3 subset.

We suggest you remove the books3 data from this dataset

In the meantime switching the dataset to private while you remove the infringing data might be a good idea

Thanks,

HF moderation team



Hugging Face

 **Datasets:**  P1ayer-1/**books-3**  like 5 **Dataset card** Files Community **1** **Access to this dataset has been disabled**

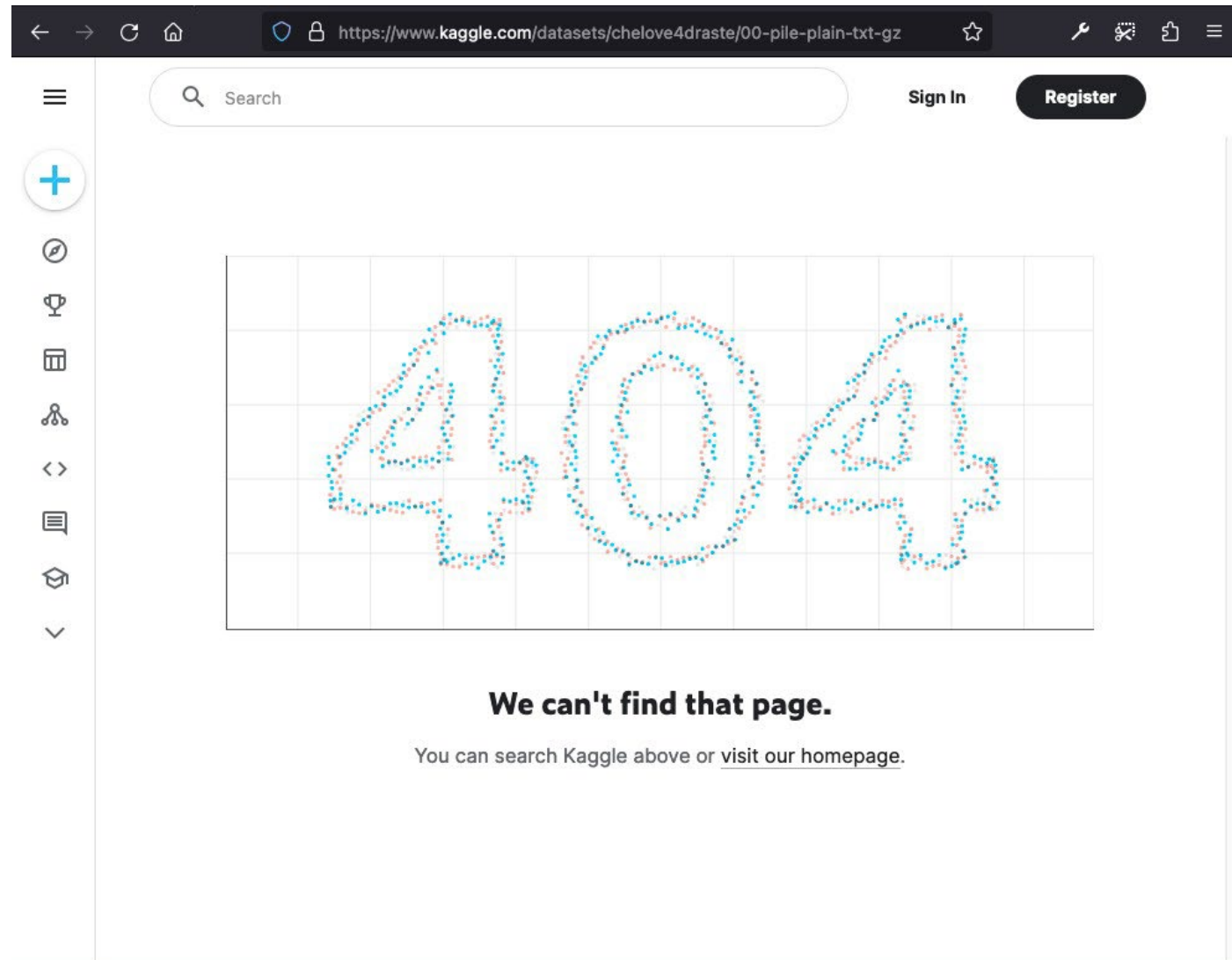
disabled due to <https://huggingface.co/datasets/P1ayer-1/books-3/discussions/1>

Books3 - takedown procedure



A screenshot of the Kaggle dataset page for "00_pile_plain_txt". The browser address bar shows the URL: <https://www.kaggle.com/datasets/chelove4draste/00-pile-plain-txt>. The page features a search bar, "Sign In" and "Register" buttons, and a "Download (15 GB)" button. The dataset title "00_pile_plain_txt" is prominently displayed, along with a "Data Card" tab and a "New Notebook" button. Below the title, there is an "About Dataset" section with the text "No description available" and a "Usability" score of 1.25. Other metadata includes "License: Unknown" and "Expected update frequency: Not specified". A purple abstract image is visible on the right side of the page.

A screenshot of the Kaggle dataset page for "00_pile_plain_txt", showing the content of the "output.txt" file. The browser address bar shows the URL: <https://www.kaggle.com/datasets/chelove4draste/00-pile-plain-txt>. The page title is "00_pile_plain_txt" and the file size is "44.08 GB". The content of the file is displayed in a monospaced font, starting with "Working over the theme was probably one of the hardest tasks I had to face." and continuing with a detailed description of a game development process. The text includes phrases like "Originally, I had an idea of what kind of game I wanted to develop, gameplay wise...", "In the end, the problem with a theme like 'Evolution' in a game is that evolution...", "In a game, you need to control something to reach an objective. That control goes...", "Hence, my biggest dilemma when deciding what to create was not with what I wanted...", "This is a problem, of course, every other contestant also had to face. And judging...", "Alas, this is just a fun competition and after a while I decided not to be as strict...", "My initial idea was to create something where humanity tried to evolve to a next level...", "Borgs were my next inspiration, as their whole hypothesis fit pretty well into the...", "The third and final idea came to me through my girlfriend, who somehow gave me the...", "Conversations with my inspiring co-worker Roushey (who also created the 'Mechanical...', "So the idea evolved more or less into this: you are sitting a table. You have your...", "Your plate can spawn little pieces of pasta. You do so by 'ordering' them through...", "Once spawned, your pastas start flying around. Their instinct is to fly to other plates...", "Your pasta doesn't like other people's pasta, so if they meet, they shoot sauce at...", "Once a pasta is in the vicinity of a plate, it starts conquering it for its team....", "You get points every second for every plate you own.", "Over time, the concept also evolved to use an Italian bistro as its main scenario.", "Carlos, Carlos' Bistro's founder and owner", "Setup", "No major changes were made from my work setup. I used FDT and Starling creating an...", "One big change for me was that I livestreamed my work through a twitch.tv account.", "Knowing the video was being recorded also made me a lot more self-conscious about...", "My own stream was probably boring to watch since I was coding for the most time. B...", "Summary", "1 file", ".txt", "1"




The image shows a browser window displaying a 404 error page on Kaggle. The address bar shows the URL `https://www.kaggle.com/datasets/chelove4draste/00-pile-plain-txt-gz`. The page features a search bar, "Sign In" and "Register" buttons, and a vertical navigation menu on the left. The main content area contains a large "404" rendered in a dotted font on a grid background. Below the grid, the text reads: "We can't find that page. You can search Kaggle above or [visit our homepage.](#)"

Books3 - takedown procedure



← Post


 **Shawn Presser** ✓
@theshawwn

Update: You can download books3 directly from this URL:
dl.books3.is/books3.tar.gz (36.8 GiB, takes ~15min)

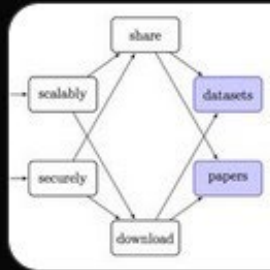
And you can download books1 (bookcorpus) from here:
dl.books3.is/books1.tar.gz (2.2 GiB, takes ~2min)

8:48 AM · Aug 10, 2023 · 781 Views

1 Repost 1 Quote 10 Likes 8 Bookmarks

 **Shawn Presser** ✓ @theshawwn · Aug 19

For the moment, the method to acquire books3 is to go to academictorrents.com, search The Pile, and use a torrent client that can download specifically the books3 part (37GB). Unless you have 800GB of space and don't mind waiting a day or so for everything.



```
graph TD; share[share] --> scalably[scalably]; share --> datasets[datasets]; scalably --> download[download]; datasets --> download; securely[securely] --> download; papers[papers] --> download;
```

academictorrents.com
Academic Torrents
A distributed system for sharing enormous datasets - for researchers, by researchers. The result is a ...

2 2 26 2,281

Books3 - disruption of distribution



This Tweet is unavailable. [Learn more](#)



Books3 - disruption of distribution



Academic Torrents

https://academicorrents.com/details/0d366035664fdf51cfbe9f733953ba325776e667

paper, author, or dataset Search

The Pile An 800GB Dataset of Diverse Text for Language Modeling

EleutherAI

Home Technical 0/0 Comments 0 Collections 0

A DMCA notice has been issued for this torrent

Date	2023-08-24 13:54:33
Submitter Name	Thomas Heldrup, Vesterbrogade 15, 1 floor, 1620 Copenhagen V, Denmark
Submitter Email	Thomas.heldrup@rettighedsalliancen.dk
Provide a description of the content in question:	"The book "Afrikas Horn" by Wilbur Smith, published by Lindhardt og Ringhof A/S in Denmark. There are additional 108 works we represent that are infringed on the URL. On the following link you can see an official description of "Afrikas Horn" by Wilbur Smith: https://www.lindhardtogringhof.dk/afrikas-horn-3 "
How are you authorized to make the request?	Authorised agent
How is the content not covered under the Fair Use Act sections 107 or 108?	The work originates from an illegal filesharing site called bibliotik.me (this explicit from the paper documenting "The Pile" found here: https://arxiv.org/abs/2101.00027). As the origin of the copy of the content is an illegal source the content cannot be claimed to fall under the Fair Use doctrine.
Provide a statement that the complaining party has a good faith belief.	I have good faith that the use of the work in this notice is not authorised by the copyright owner its agent, or the law.

EleutherAI ThePile v1 (51 files)

Build secure applications

Books3 - AI model creators



LLaMa 1

FEB 23

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Gutenberg and Books3 [4.5%]. We include two book corpora in our training dataset: the Gutenberg Project, which contains books that are in the public domain, and the Books3 section of ThePile (Gao et al., 2020), a publicly available dataset for training large language models. We perform deduplication at the book level, removing books with more than 90% content overlap.



No transparency

LLaMa 2

JULY 23

2.1 Pretraining Data

Our training corpus includes a **new mix of data** from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.

StableLM-Alpha

APR 23

☰ README.md

StableLM-Alpha

StableLM-Alpha models are trained on a new dataset that builds on [The Pile](#), which contains 1.5 trillion tokens, roughly 3x the size of The Pile. The context length for these models is 4096 tokens.

As a proof-of-concept, we also fine-tuned the model with [Stanford Alpaca](#)'s procedure using a combination of five recent datasets for conversational agents: Stanford's [Alpaca](#), Nomic-AI's [gpt4all](#), RyokoAI's [ShareGPT52K](#) datasets, Databricks labs' [Dolly](#), and Anthropic's [HH](#). We will be releasing these models as StableLM-Tuned-Alpha.

Size	StableLM-Base-Alpha	StableLM-Tuned-Alpha	Training Tokens	Parameters	Web Demc
3B	checkpoint	checkpoint	800B	3,638,525,952	
7B	checkpoint	checkpoint	800B	7,869,358,080	Hugging Face

StableLM-3B

OCT 23

StableLM-3B-4E1T

Technical report for StableLM-3B-4E1T

[Jonathan Tow](#), [Marco Bellagente](#), [Dakota Mahan](#), [Carlos Riquelme Ruiz](#)

[Model Architecture](#)

[Training Data](#)

[Training Procedure](#)

[Downstream Results](#)

[System Details](#)

[Conclusion](#)

[Acknowledgments](#)

[References](#)

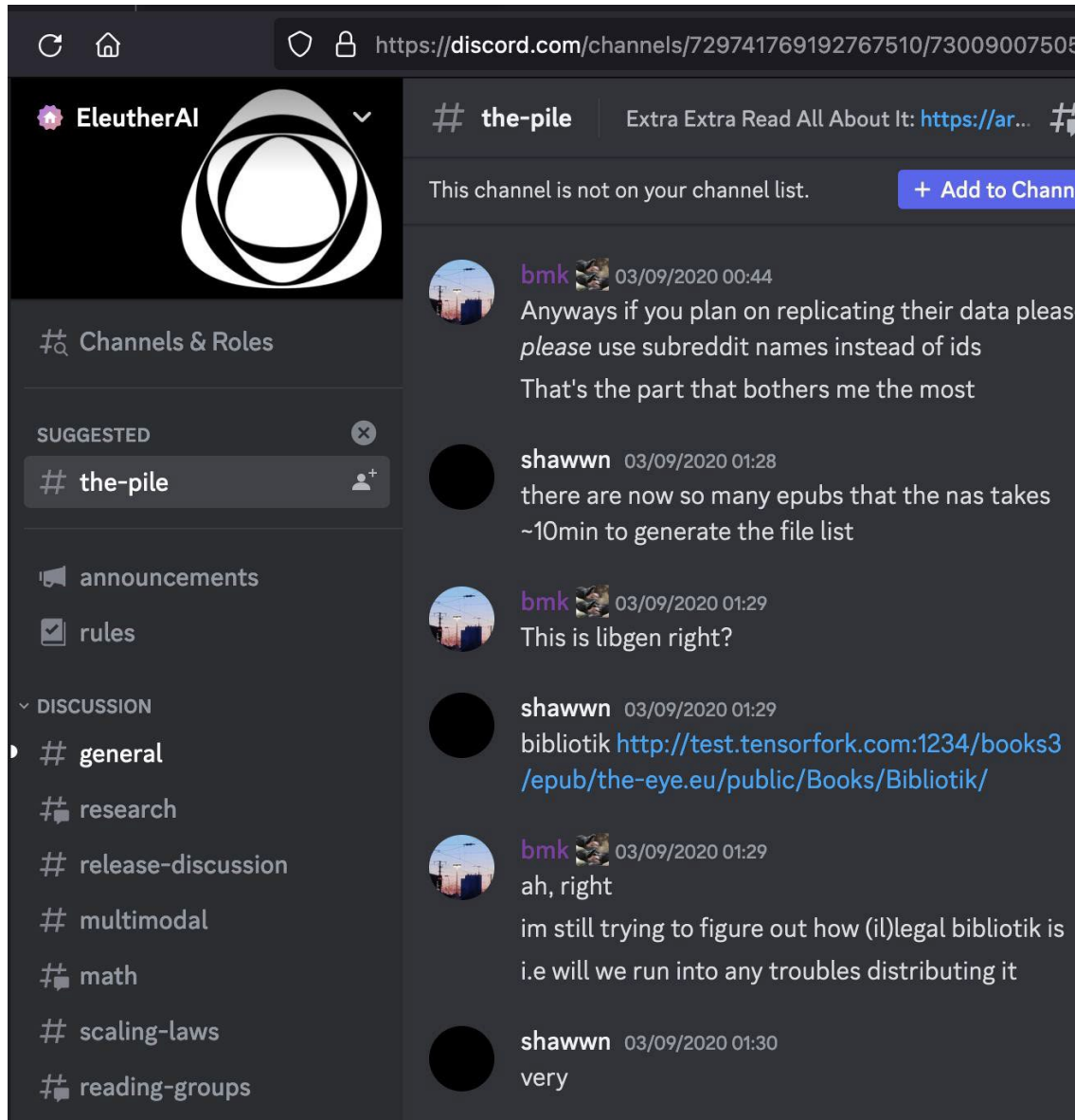
▾ Training Data

The dataset is comprised of a filtered mixture of open-source large-scale datasets available on the [HuggingFace Hub](#): Falcon RefinedWeb extract ([Penedo et al., 2023](#)), RedPajama-Data ([Together Computer, 2023](#)) and The Pile ([Gao et al., 2020](#)), both without the [Books3 subset](#), and StarCoder ([Li et al., 2023](#)). The complete list is provided in Table 1.



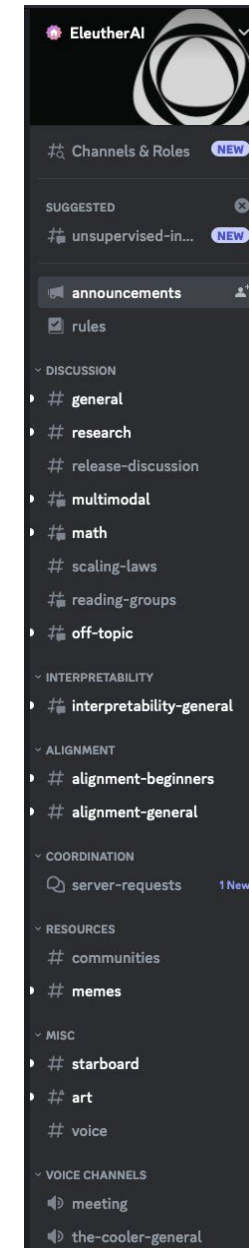
EleutherAI Discord

Screenshot JULY 23



Deleted # the-pile

Screenshot SEP 23



Bloomberg

Dear Mr. Heldrup,

Thanks for your email. As the paper we published in March indicated, we used a number of different data sources to train the initial BloombergGPT model, which was developed for research purposes. As we continue to research and develop BloombergGPT, we are evaluating what datasets to include when training future versions of the model for use in a commercial capacity. We will not include the Books3 dataset among the data sources used to train future versions of BloombergGPT.

Please let me know if you have any questions.

Paul Ramundo
Bloomberg Legal Department
Bloomberg L.P.

Books3 - concluding thoughts

Illegal copies of content appear in publicly available datasets.

Current enforcement tools such as takedown notices can be used to remove illegal dataset from major dataset hosting platforms.

Takedown process is currently slow at Kaggle and not fully streamlined at Hugging Face yet with much leeway given to uploaders.

There are technical barriers to discovery of illegal datasets and it requires many resources to maintain.



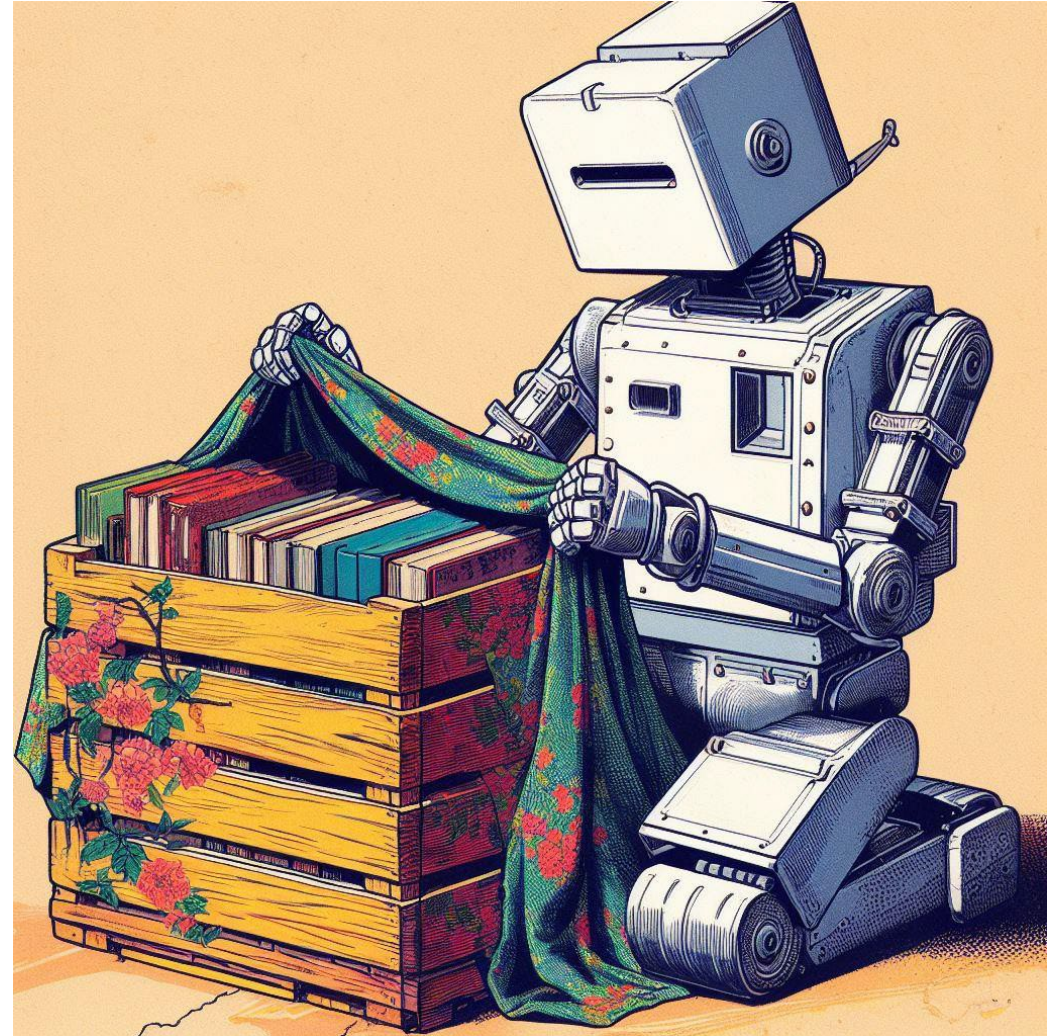
Created by Thomas Heldrup with Bing Image Creator

Books3 - concluding thoughts

The Books3 case is unique for the level of transparency at all the key places e.g. works & source.

It shows that transparency is crucial to enabling rightsholders to enforce illegal use of their work.

AI model creators and providers are willing to use illegal copies of works to train AI and they won't be transparent moving forward.



Created by Thomas Heldrup with Bing Image Creator