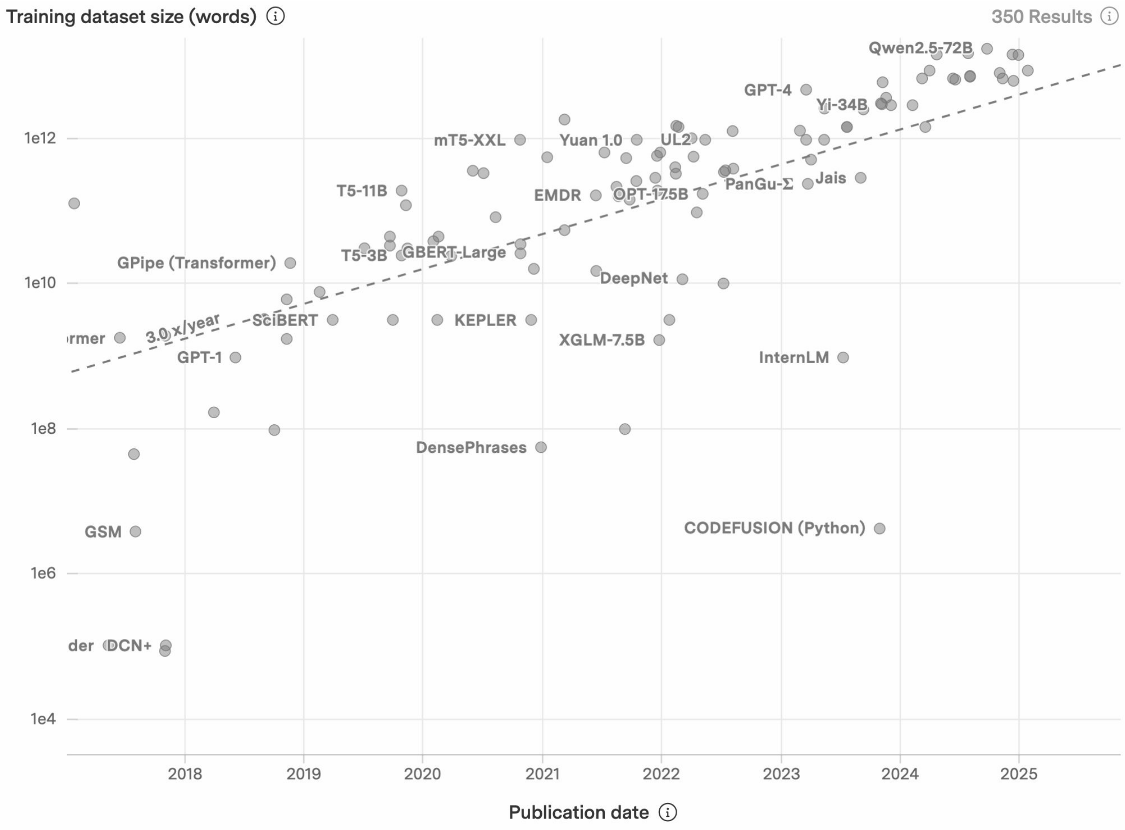
The Common Pile

Brian Lester
Big Data Value Association
2025/06/27

UNIVERSITY OF
TORONTO

Training datasets are growing



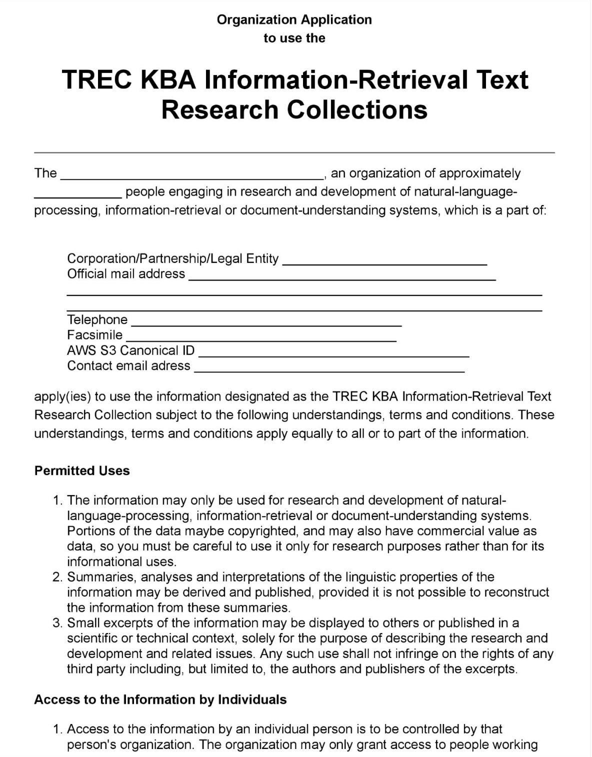
"Data on Notable AI Models" by EpochAI

How are people getting this data?

Days gone by

- Small Datasets
- More interested in the annotations than the text
- 💸💸💸 and ULAs

```
SOCCER   NN  I-NP O
-        :   O    O
JAPAN    NNP I-NP I-LOC
GET      VB  I-VP O
LUCKY    NNP I-NP O
WIN      NNP I-NP O
,        ,   O    O
CHINA    NNP I-NP I-PER
IN       IN  I-PP O
SURPRISE DT  I-NP O
DEFEAT   NN  I-NP O
.        .   O    O
```



Organization Application
to use the

TREC KBA Information-Retrieval Text Research Collections

The ______________________________, an organization of approximately __________ people engaging in research and development of natural-language-processing, information-retrieval or document-understanding systems, which is a part of:

Corporation/Partnership/Legal Entity ______________________
Official mail address ______________________________
____
____
Telephone ______________________________
Facsimile ______________________________
AWS S3 Canonical ID ______________________________
Contact email adress ______________________________

apply(ies) to use the information designated as the TREC KBA Information-Retrieval Text Research Collection subject to the following understandings, terms and conditions. These understandings, terms and conditions apply equally to all or to part of the information.

Permitted Uses

1. The information may only be used for research and development of natural-language-processing, information-retrieval or document-understanding systems. Portions of the data maybe copyrighted, and may also have commercial value as data, so you must be careful to use it only for research purposes rather than for its informational uses.
2. Summaries, analyses and interpretations of the linguistic properties of the information may be derived and published, provided it is not possible to reconstruct the information from these summaries.
3. Small excerpts of the information may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development and related issues. Any such use shall not infringe on the rights of any third party including, but limited to, the authors and publishers of the excerpts.

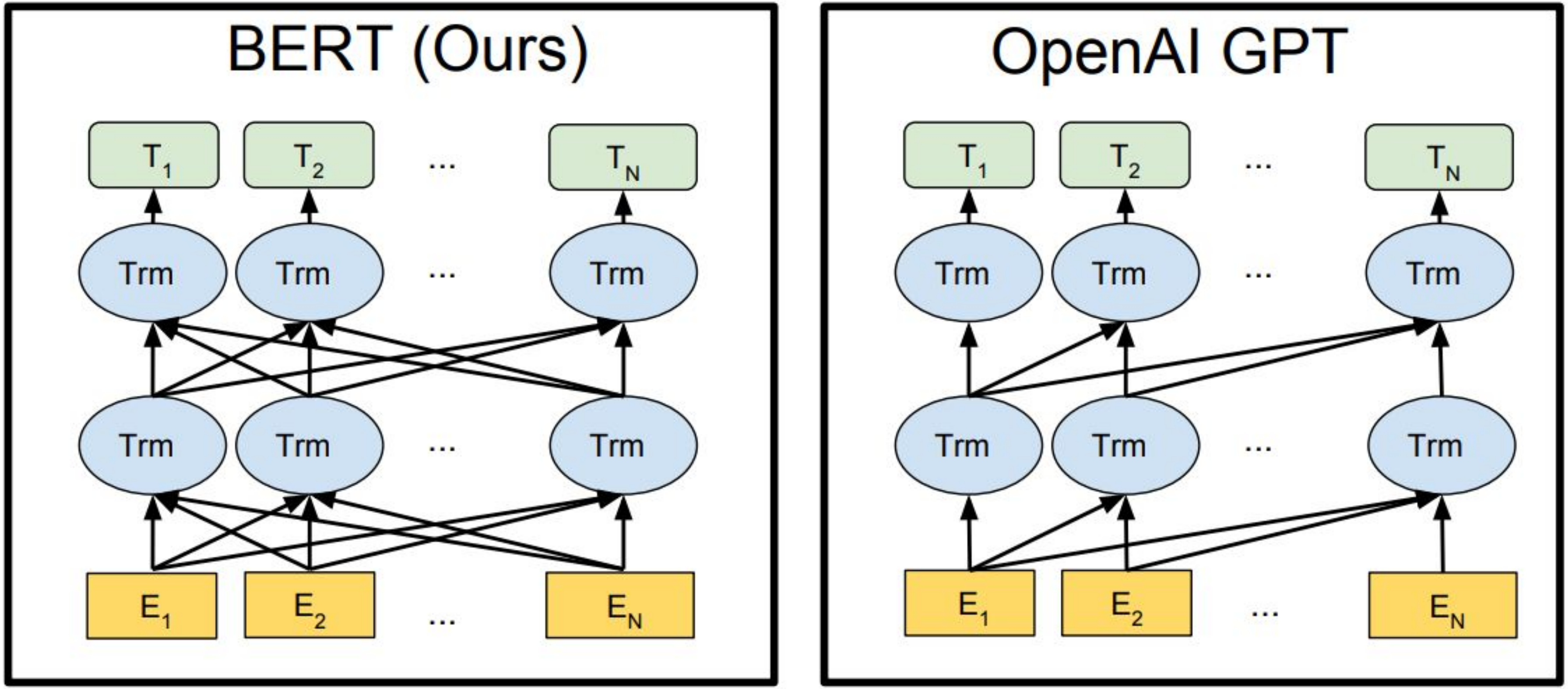
Access to the Information by Individuals

1. Access to the information by an individual person is to be controlled by that person's organization. The organization may only grant access to people working

Linguistic Data Consortium

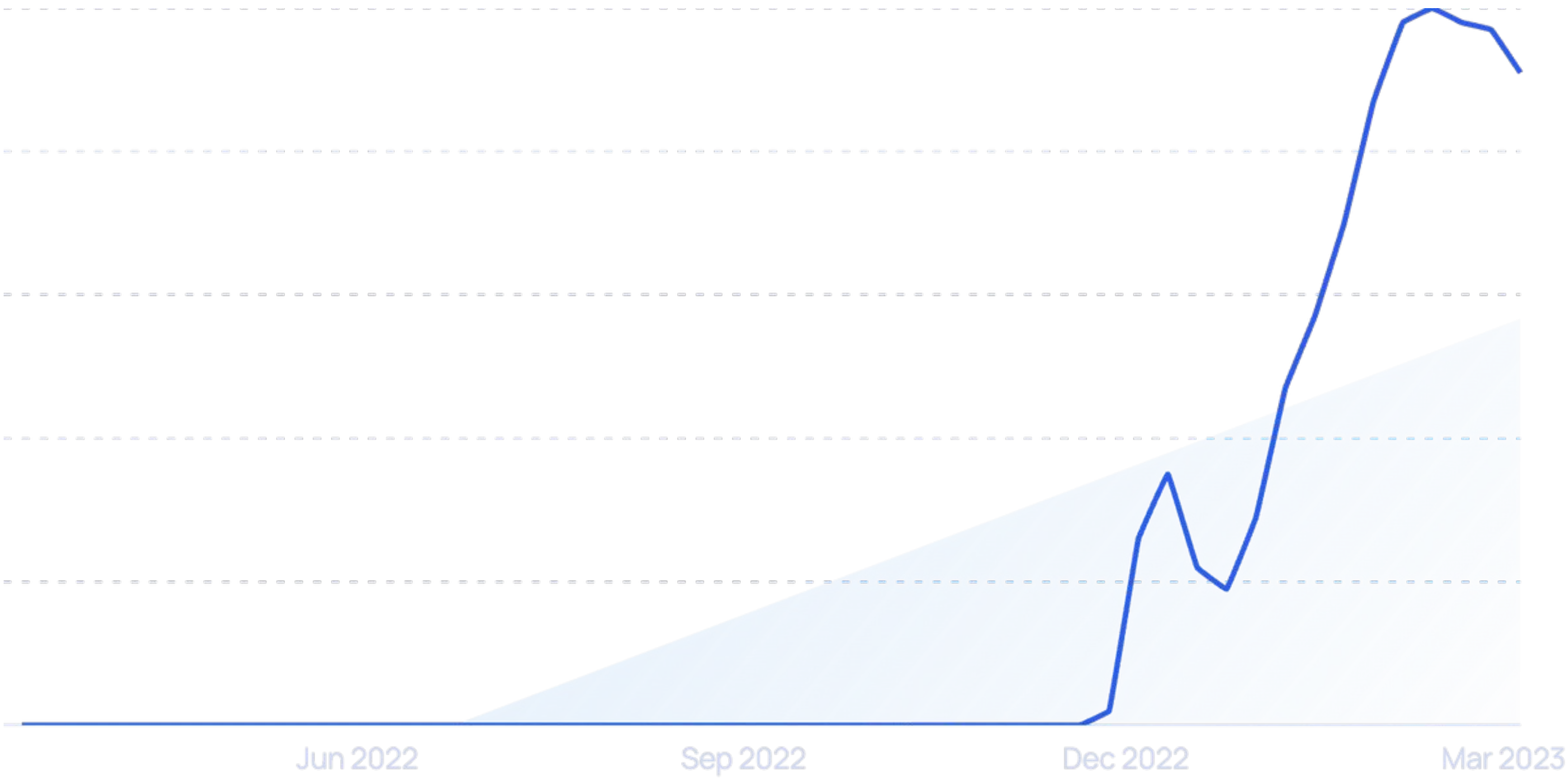
https://www.ldc.upenn.edu/
https://trec.nist.gov/
https://ic2s2-2024.org/
Conll 2003 Shared Task

Transfer learning and pretraining



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Devlin et. al.

Scraping webtext



https://explodingtopics.com/blog/chatgpt-users

People are upset… and litigious

Coders

1. Joseph Saveri Firm: overview, complaint

Writers

2. Joseph Saveri Firm: overview, complaint
3. Authors Guild & Alter: overview, complaint
4. Nicholas Gage: overview & complaint

YouTubers

5. Millette: overview, complaint

Media

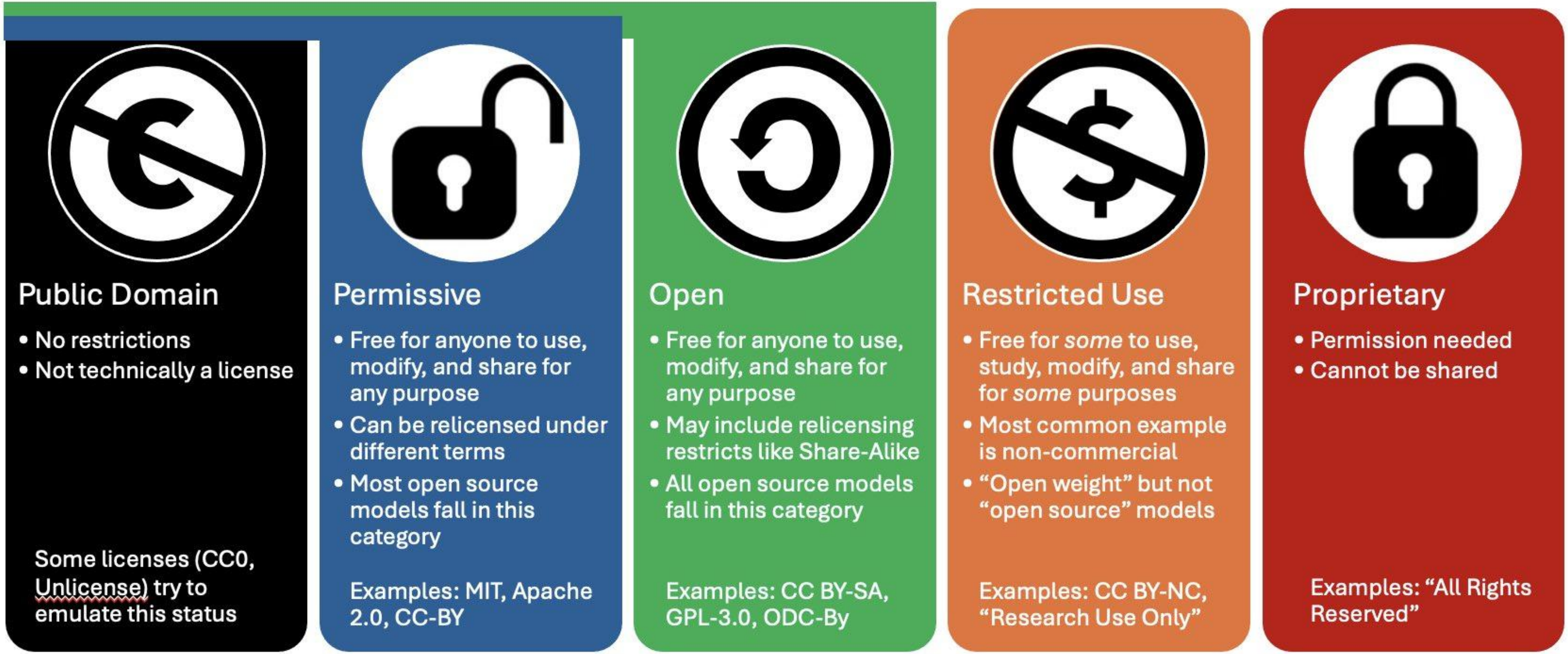
6. New York Times: overview, complaint
7. Intercept Media: overview, complaint
8. Raw Story & Alternet: overview, complaint
9. Denver Post & seven others: overview, complaint
10. Center for Investigative Reporting: overview, complaint

https://forum.effectivealtruism.org/posts/EdBkBBFkCaHrw2iwL/twelve-lawsuits-against-openai

Is proprietary data needed for LLMs?

> *“… it would be impossible to train today’s leading AI models without using copyrighted materials.”*
>
> — OpenAI

What is Openly Licensed Text?



Public Domain

- No restrictions
- Not technically a license

Some licenses (CC0, Unlicense) try to emulate this status

Permissive

- Free for anyone to use, modify, and share for any purpose
- Can be relicensed under different terms
- Most open source models fall in this category

Examples: MIT, Apache 2.0, CC-BY

Open

- Free for anyone to use, modify, and share for any purpose
- May include relicensing restricts like Share-Alike
- All open source models fall in this category

Examples: CC BY-SA, GPL-3.0, ODC-By

Restricted Use

- Free for *some* to use, study, modify, and share for *some* purposes
- Most common example is non-commercial
- "Open weight" but not "open source" models

Examples: CC BY-NC, "Research Use Only"

Proprietary

- Permission needed
- Cannot be shared

Examples: "All Rights Reserved"

https://x.com/AiEleuther/status/1931021652772462732

https://opendefinition.org/

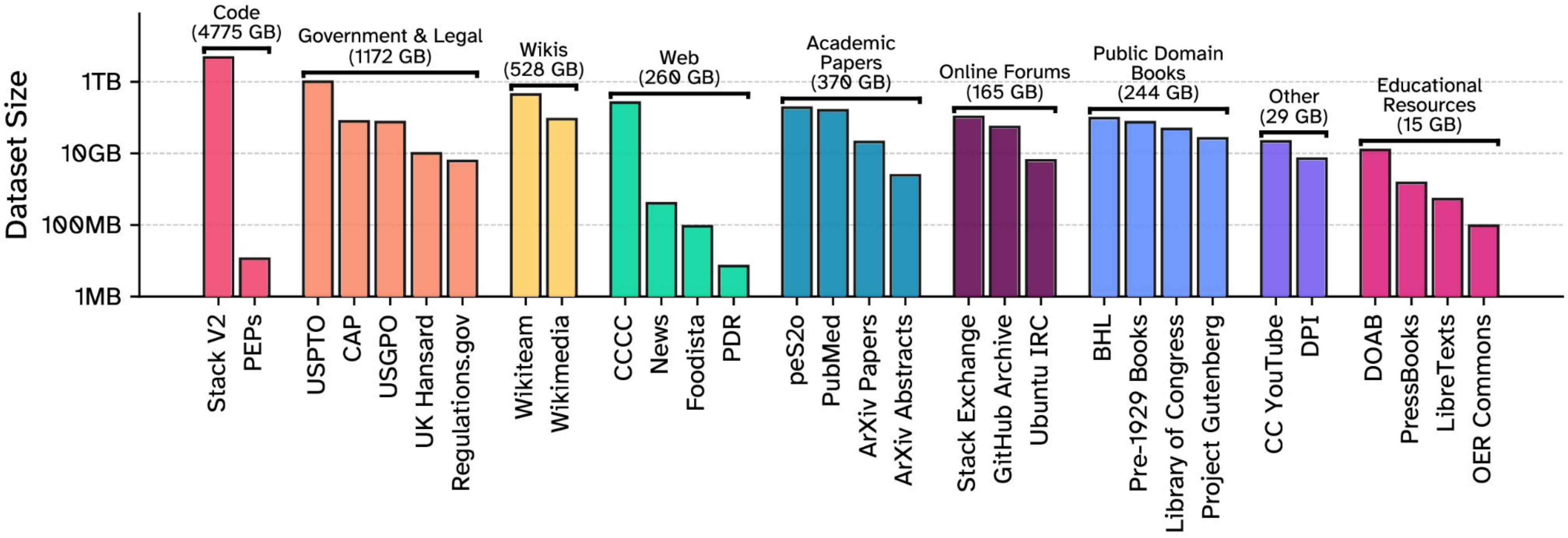
So… which ones?

- Public Domain/CC0 ✅
  - Old Books
  - Government text
- Creative Commons
  - Attribution Required (CC BY, CC BY-SA) ✅
  - Non-commercial is out (CC NC) ❌
  - Does “No derivative works” include models? (CC ND) ❌
    - Do models need attribution? 🤷‍♀️
- Blue Oak Council Licenses ✅
  - BSD, MIT, Apache 2.0, etc.

- https://github.com/r-three/common-pile/blob/main/common_pile/licenses.py

The Common Pile Raw Source Sizes

- 8TB total



Examples—Mediawiki Wikis

```
Information theory

Information theory is the mathematical study of the quantification, storage, and communication of
information. The field was originally established by the works of Harry Nyquist and Ralph Hartley, in
the 1920s, and Claude Shannon in the 1940s. The field, in applied mathematics, is at the intersection
of probability theory, statistics, computer science, statistical mechanics, information engineering,
and electrical engineering.

A key measure in information theory is entropy. Entropy quantifies the amount of uncertainty involved
in the value of a random variable or the outcome of a random process. For example, identifying the
outcome of a fair coin flip (with two equally likely outcomes) provides less information (lower
entropy, less uncertainty) than specifying the outcome from a roll of a die (with six equally likely
outcomes). Some other important measures in information theory are mutual information, channel
capacity, error exponents, and relative entropy. Important sub-fields of information theory include
source coding, algorithmic complexity theory, algorithmic information theory and
information-theoretic security.

Applications of fundamental topics of information theory include source coding/data compression (e.g.
for ZIP files), and channel coding/error detection and correction (e.g. for DSL). Its impact has been
crucial to the success of the Voyager missions to deep space, the invention of the compact disc, the
feasibility of mobile phones and the development of the Internet. The theory has also found
applications in other areas, including statistical inference, cryptography, neurobiology, perception,
linguistics, the evolution and function of molecular codes (bioinformatics), thermal physics, …
```

Examples—Wikiteam Wikis

Man Vs. Society
Katniss’s Society is riddled with problems. This is the exact reason why she is forced to enter the hunger games. The hunger Games serve as a reminder that the Capitol is in control. This is the most important conflict in the story.

Man Vs Man
Katniss is put in a situation where she is forced to fight and kill for her life a gainst 23 other tributes untill she is the only survivor left. She is not able to trust or count on any of the other tributes, for they might just be planning to kill her. The Hunger Games is all about the fights, the violence, and the deaths of these tributes as they struggle to return home alive.

Man Vs. Nature
Katniss is faced with many tough situations where she has to fend for herself in nature. When she is forced into the hunger games she must find water on her own, along with hunting food food and nutrience in order to survive. It is clear that nature is not working with Katniss, it is working against her.

Man Vs. Self
She also fights to find hope in a hopeless world where everyone is out to get her. Another internal battle she faces is to keep her humanity despite the atrocities she’s seen and even committed.

Examples—UK Hansard

Department for Business, Innovation and Skills

Burberry: North Korea

Lord Alton of Liverpool: To ask Her Majesty’s Government what assessment they have made of reports that North Korean workers are producing garments for Burberry Group Inc.

Baroness Neville-Rolfe: We have made no such assessment. The Modern Slavery Act, enacted last year, includes a demanding transparency in supply chains disclosure provision. This provision requires businesses operating in the UK, with an annual turnover of more than £36 million, to produce a slavery and human trafficking statement setting out the steps they have taken during the financial year to ensure that slavery and human trafficking is not taking place in any of its supply chains or its own business. The provision obliges eligible companies to publish this information on their website, with a conspicuous link from their homepage, so that consumers, investors and the general public know what steps businesses are taking in this regard.

Students: Loans

Lord Myners: To ask Her Majesty’s Government what assessment they have made of the impact on graduates of freezing the income level at which student loans become repayable, and what estimate they have made of the number of students who would be rendered liable...

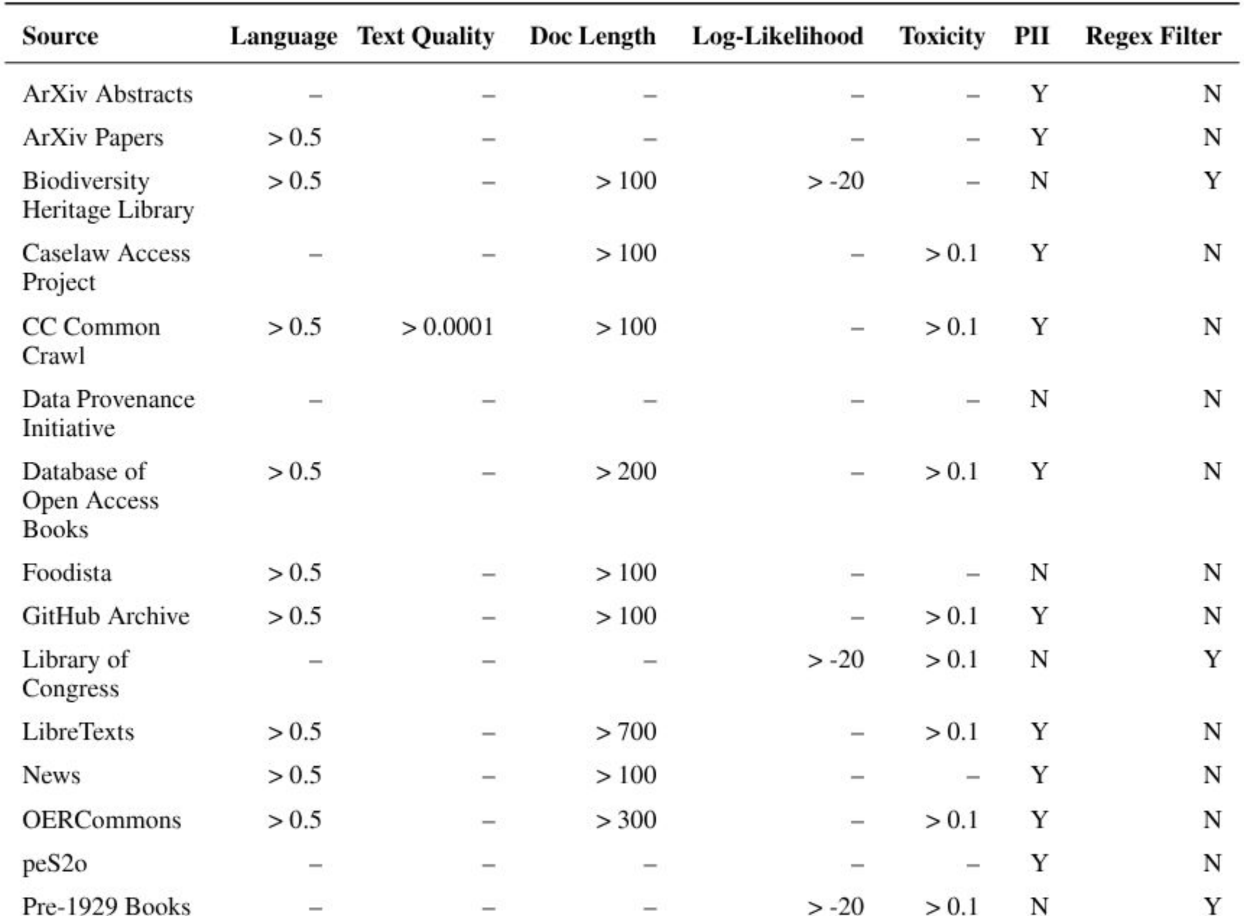
Examples—Github Archives

```
trainer class is None when training with my own data
how to solve this problem?

Please make sure the trainer file exists in the path:
nnFormer/training/network_training/

Another possible reason is that the trainer file exists in the above path, but the Class name in the
trainer is not same with the trainer file

do you run nnformer on your own dataset successfully?

what does that mean？I use my own data only have foreground and background
I ran python inference_synapse.py
but I got this

open the dice_pre.txt
I got

Is the number of classs not set properly
so what should I do
```

Examples—PEPs

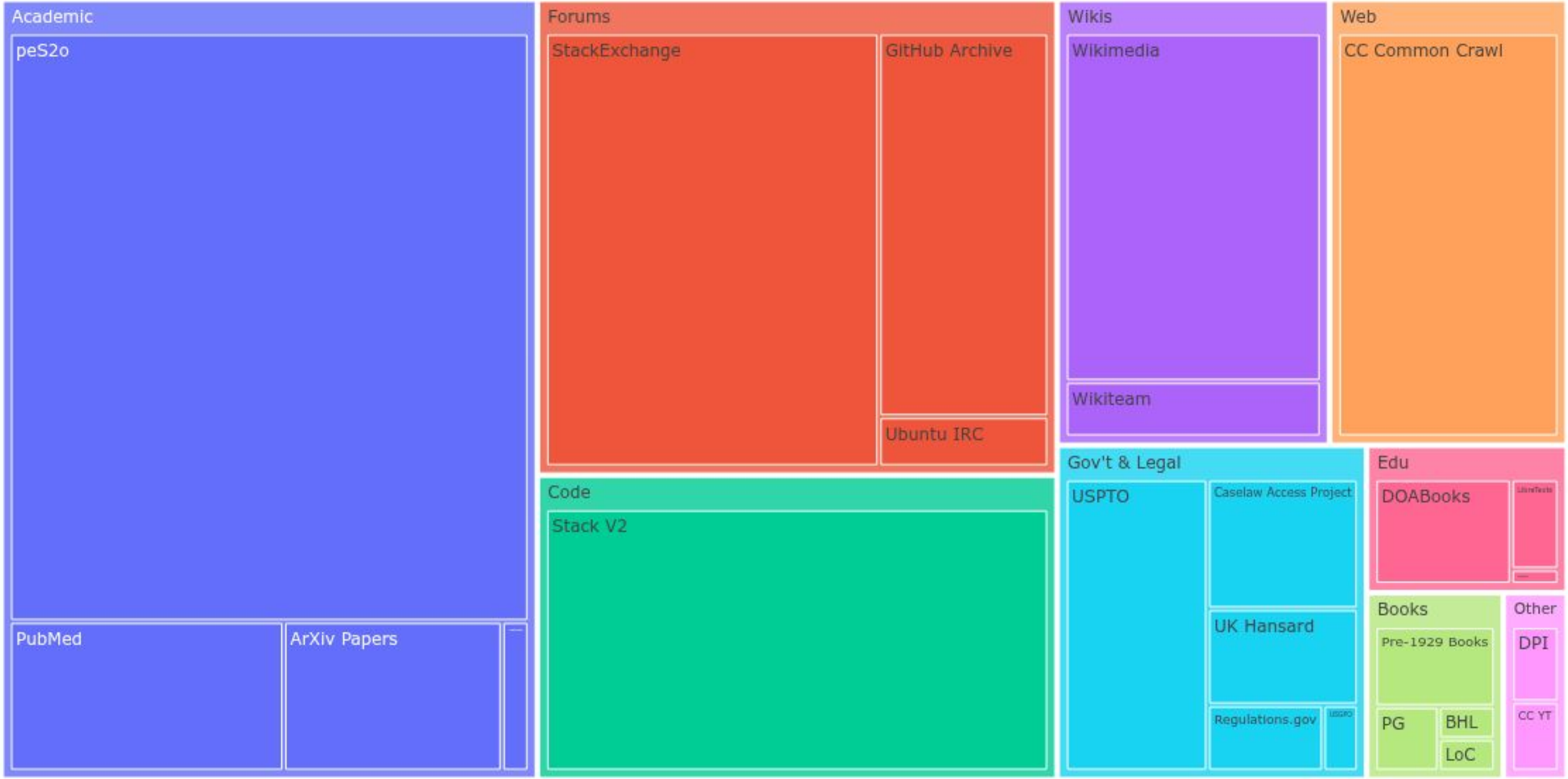
```
PEP: 3144 Title: IP Address Manipulation Library for the Python Standard
Author: Peter Moody <pmoody@google.com>
BDFL-Delegate: Alyssa Coghlan
Discussions-To: ipaddr-py-dev@googlegroups.com
Status: Final
Type: Standards Track
Content-Type: text/x-rst
Created: 06-Feb-2012 Python-Version: 3.3
Resolution: https://mail.python.org/pipermail/python-dev/2012-May/119474.html

Abstract

This PEP proposes a design and for an IP address manipulation module for
python.

PEP Acceptance

This PEP was accepted by Alyssa Coghlan on the 15th of May, 2012.

Motivation

Several very good IP address modules for python already exist...
```

Raw Text != A Training Dataset

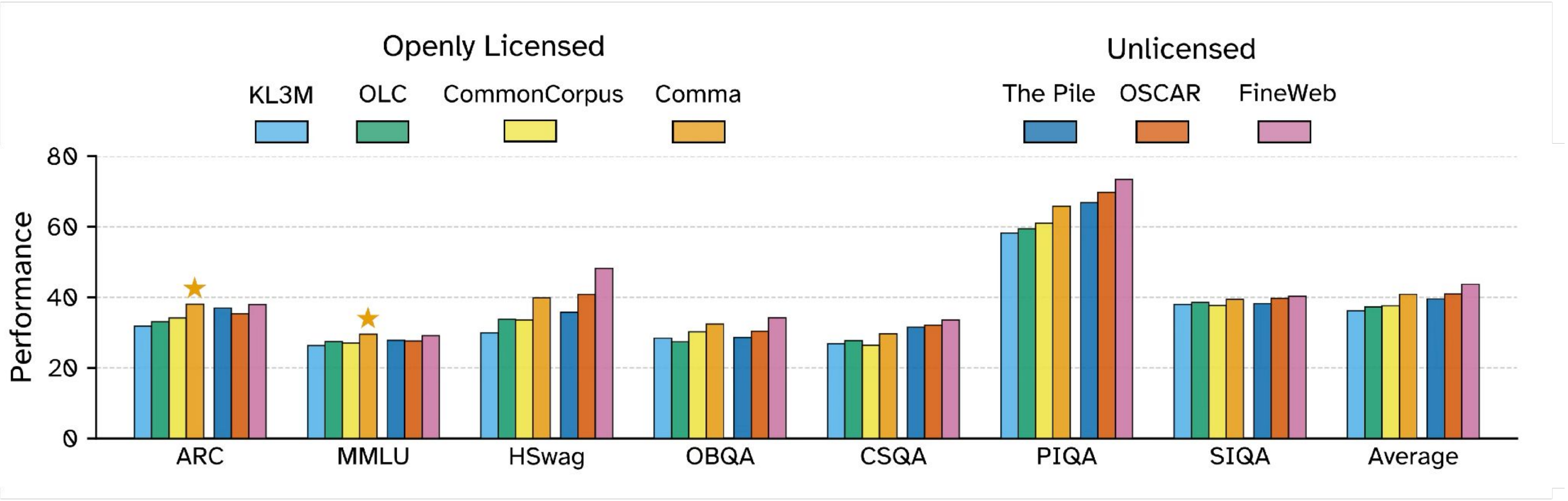


Source	Language	Text Quality	Doc Length	Log-Likelihood	Toxicity	PII	Regex Filter
ArXiv Abstracts	–	–	–	–	–	Y	N
ArXiv Papers	> 0.5	–	–	–	–	Y	N
Biodiversity Heritage Library	> 0.5	–	> 100	> -20	–	N	Y
Caselaw Access Project	–	–	> 100	–	> 0.1	Y	N
CC Common Crawl	> 0.5	> 0.0001	> 100	–	> 0.1	Y	N
Data Provenance Initiative	–	–	–	–	–	N	N
Database of Open Access Books	> 0.5	–	> 200	–	> 0.1	Y	N
Foodista	> 0.5	–	> 100	–	–	N	N
GitHub Archive	> 0.5	–	> 100	–	> 0.1	Y	N
Library of Congress	–	–	–	> -20	> 0.1	N	Y
LibreTexts	> 0.5	–	> 700	–	> 0.1	Y	N
News	> 0.5	–	> 100	–	–	Y	N
OERCommons	> 0.5	–	> 300	–	> 0.1	Y	N
peS2o	–	–	–	–	–	Y	N
Pre-1929 Books	–	–	–	> -20	> 0.1	N	Y

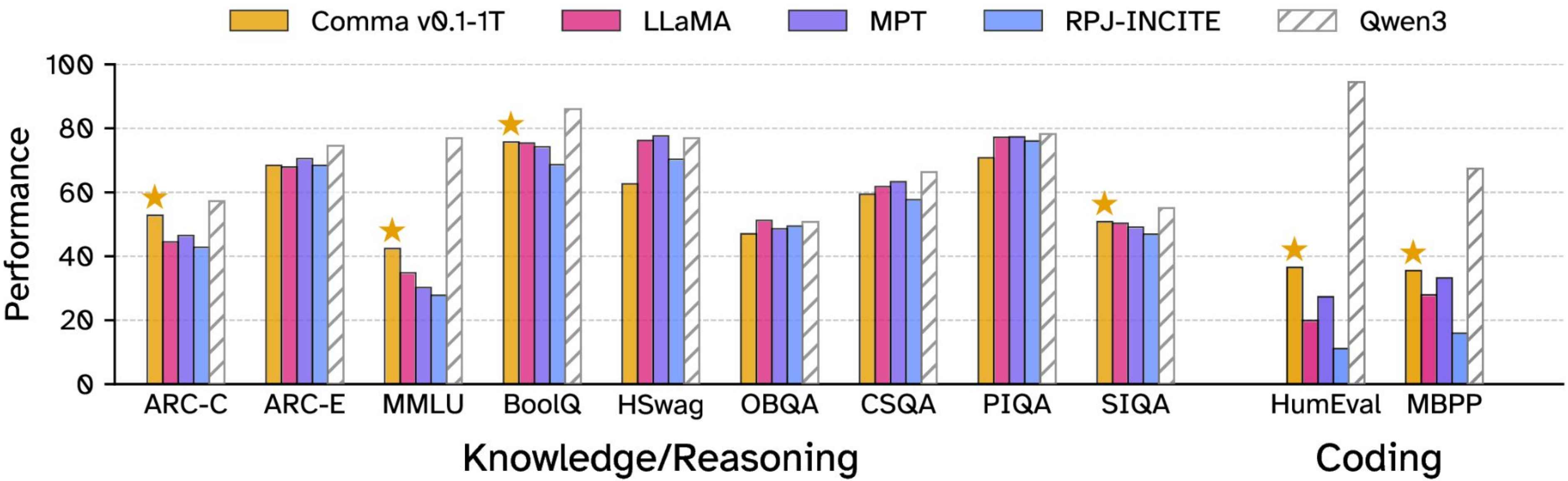
What are you made of?



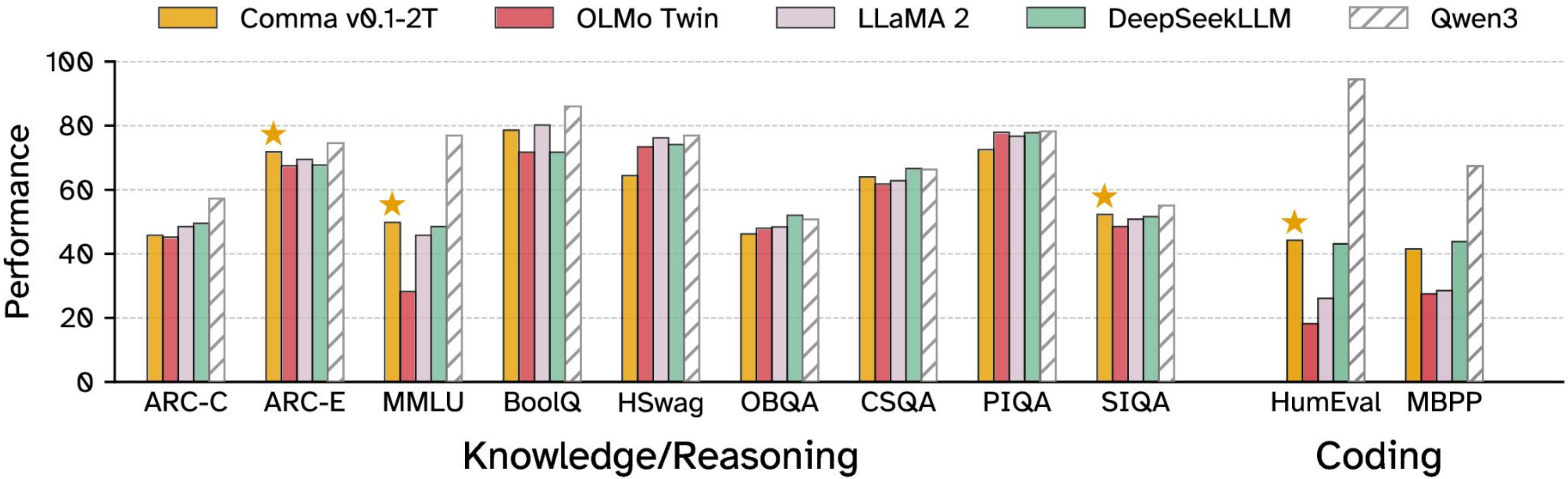
Comma v0.1


Openly Licensed
KL3M
OLC
CommonCorpus
Comma
Unlicensed
The Pile
OSCAR
FineWeb
Performance
0
20
40
60
80
ARC
MMLU
HSwag
OBQA
CSQA
PIQA
SIQA
Average

Comma 7B/1T


Comma v0.1-1T
LLaMA
MPT
RPJ-INCITE
Qwen3
Performance
0
20
40
60
80
100
ARC-C
ARC-E
MMLU
BoolQ
HSwag
OBQA
CSQA
PIQA
SIQA
HumEval
MBPP
Knowledge/Reasoning
Coding

Comma 7B/2T


Comma v0.1-2T
OLMo Twin
LLaMA 2
DeepSeekLLM
Qwen3
Performance
100
80
60
40
20
0
ARC-C
ARC-E
MMLU
BoolQ
HSwag
OBQA
CSQA
PIQA
SIQA
HumEval
MBPP
Knowledge/Reasoning
Coding

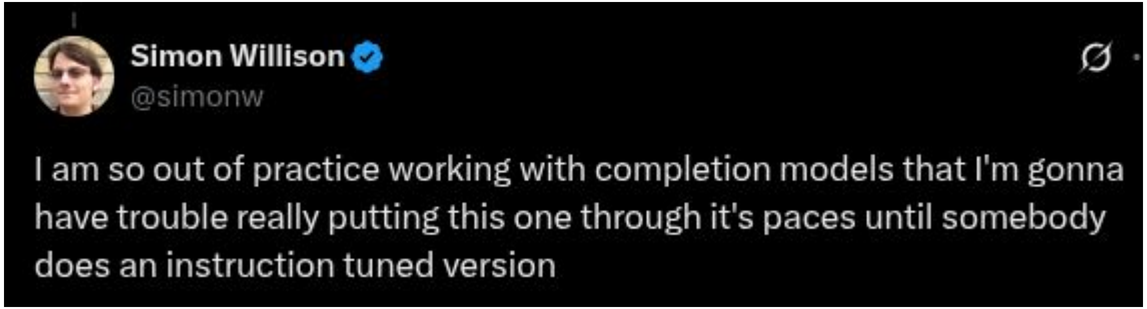
What’s Next?

What’s Next? Data Cleaning

https://gifer.com/en/2iIR
https://www.sangeministudies.info/research-projects/archaeological-ceramics-project

What’s Next? Post-Training

- Need openly licensed instruction tuning data

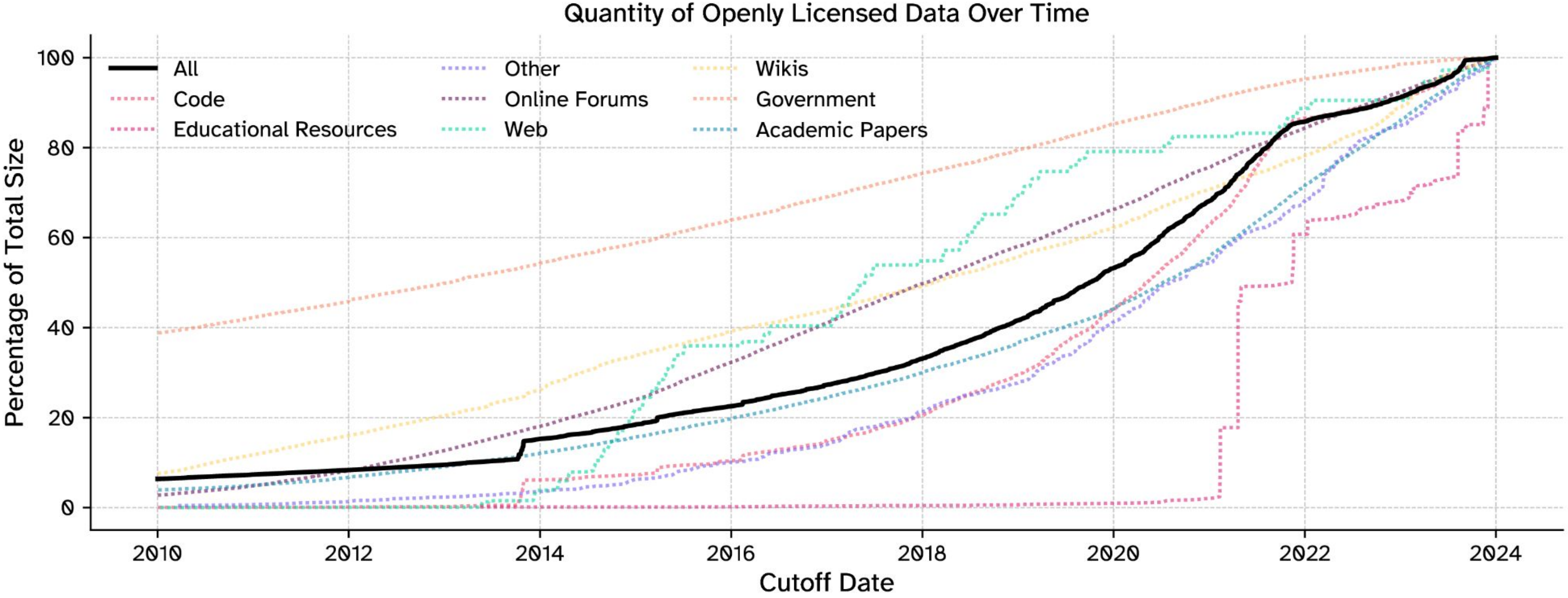


Simon Willison
@simonw

I am so out of practice working with completion models that I'm gonna have trouble really putting this one through it's paces until somebody does an instruction tuned version

https://x.com/simonw/status/1931197804137033811

What’s Next? More Data

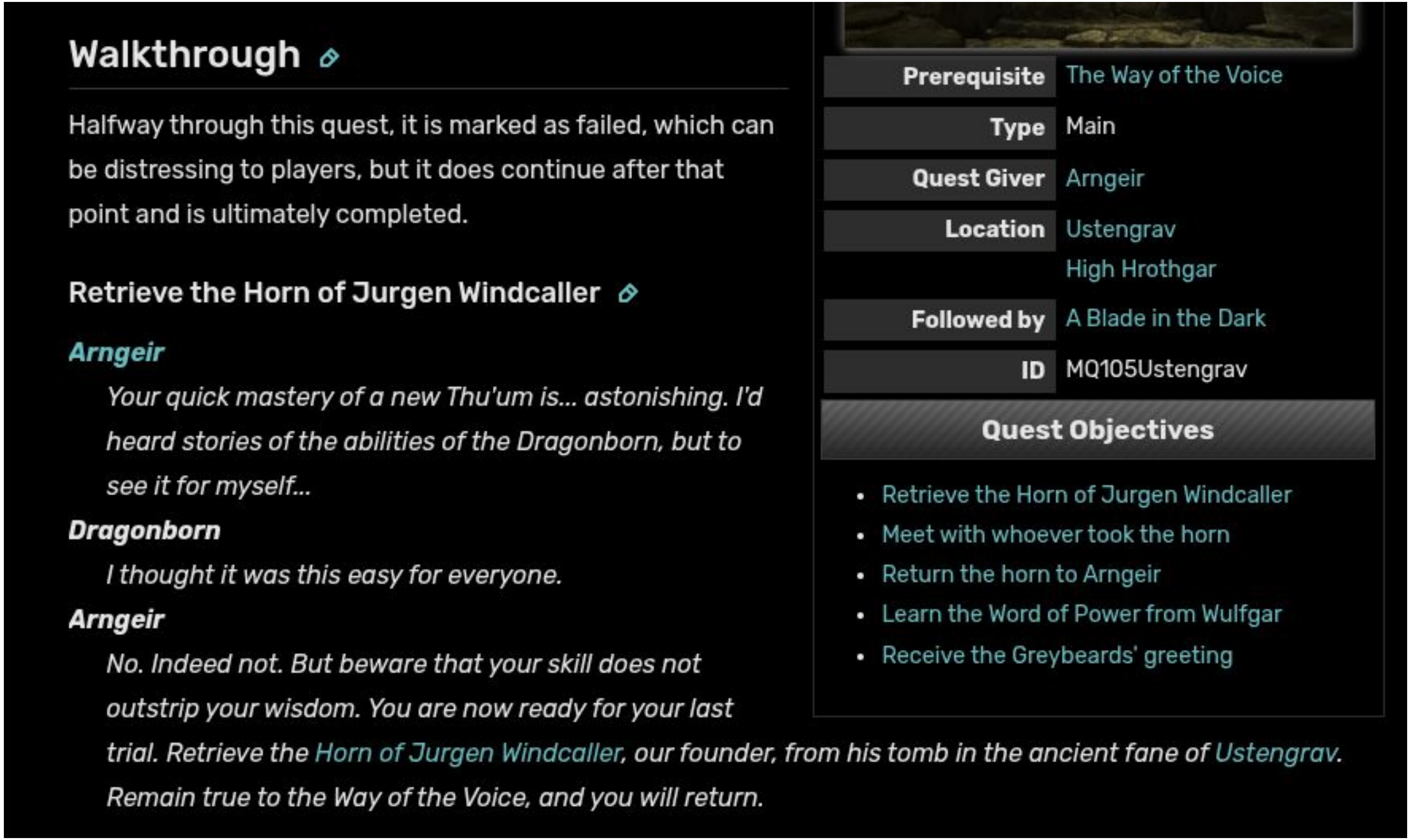


What’s Next? Synthetic Data

Applicant: AEROLINEAS EJECUTIVAS, S.A. de C.V. Date Filed: September 12, 2000 Relief requested: Exemption from 49 USC section 41301 to permit the applicant to continue to conduct passenger charter operations between Mexico and the United States, and other passenger charter operations in accordance with 14 CFR Part 212, using small equipment. If renewal, date and citation of last action: November 18, 1999; in this Docket. Applicant representative(s): Lee A. Bauer, 202-331-3300 Responsive pleadings: None.   DISPOSITION Action: Approved.   Action date: November 22, 2000 Effective dates of authority granted: November 22, 2000, through November 22, 2001. Basis for approval (bilateral agreement/reciprocity): United States-Mexico Air Transport Services Agreement of August 15, 1960, as amended and extended (Agreement). Except to the extent exempted/waived, this authority is subject to the terms, conditions, and limitations indicated: X Standard exemption conditions. Special conditions/Partial grant/Denial basis/Remarks: In the conduct of these operations, the carrier must adhere to all applicable provisions of the U.S.-Mexico Agreement. In the conduct of these operations, the carrier may only use aircraft capable of carrying no more than 60 passengers and having a maximum payload capacity of no more than 18,000 pounds (small equipment). The above grant includes authority to conduct Third and Fourth Freedom charter operations. While we have subjected, consistent with the provisions of the Agreement, Mexican carriers conducting charter operations with large aircraft to prior approval of their Third and Fourth Freedom charters (see Order 92-2-7 at 5), we determined that a Third/Fourth Freedom prior-approval requirement was not necessary on public interest grounds in the case of this carrier, since it will be conducting these operations solely with small aircraft. (Other charter operations to/from the United States under this authority, however, are subject to prior approval under 14 CFR Part 212.) Further, we are continuing to allow Mexican carriers conducting passenger charters using small equipment to make stopovers in the United States in the conduct of such operations. Action taken by: Paul L. Gretch, Director …

Caveats

Caveats—License Laundering



Walkthrough

Halfway through this quest, it is marked as failed, which can be distressing to players, but it does continue after that point and is ultimately completed.

Retrieve the Horn of Jurgen Windcaller

Arngeir

> *Your quick mastery of a new Thu'um is... astonishing. I'd heard stories of the abilities of the Dragonborn, but to see it for myself...*

Dragonborn

> *I thought it was this easy for everyone.*

Arngeir

> *No. Indeed not. But beware that your skill does not outstrip your wisdom. You are now ready for your last trial. Retrieve the Horn of Jurgen Windcaller, our founder, from his tomb in the ancient fane of Ustengrav. Remain true to the Way of the Voice, and you will return.*

Prerequisite	The Way of the Voice
Type	Main
Quest Giver	Arngeir
Location	Ustengrav High Hrothgar
Followed by	A Blade in the Dark
ID	MQ105Ustengrav

Quest Objectives

- Retrieve the Horn of Jurgen Windcaller
- Meet with whoever took the horn
- Return the horn to Arngeir
- Learn the Word of Power from Wulfgar
- Receive the Greybeards' greeting

https://skyrim.fandom.com/wiki/The_Horn_of_Jurgen_Windcaller_(quest)

Caveats—Ethics

- Would authors have used an open license if they knew GenAI was coming?
- People are developing “AI Training Allowed” licenses
- Differences between things being open/available and being easily accessible

“But the plans were on display…”
“On display? I eventually had to go down to the cellar to find them.”
“That’s the display department.”
“With a flashlight.”
“Ah, well, the lights had probably gone.”
“So had the stairs.”
“But look, you found the notice, didn’t you?”
“Yes,” said Arthur, “yes I did. It was on display in the bottom of a locked filing cabinet stuck in a disused lavatory with a sign on the door saying ‘Beware of the Leopard.”
— Douglas Adams, The Hitchhiker’s Guide to the Galaxy

Caveats—Attribution

- CC BY requires that you provide attribution (where did you get it?)
- The dataset itself clearly includes attribution
- Does the Model?
  - Does it need to?
- Can secondary systems
  - Semi-parametric Models

  provide attribution allowing for the use of proprietary data?

Conclusions

- We created a 8TB dataset of openly licensed text
    - It is the largest of its kind
    - It is the most diverse open dataset
- We trained two realistically sized models
    - Comma v0.1 7B/1T
    - Comma v0.1 7B/2T

  That are competitive with compute/data equivalent proprietary models
- **We showed that it is possible to create strong LLMs with open licensed data.**

- Dataset: https://huggingface.co/common-pile
- Model: https://huggingface.co/common-pile/comma-v0.1-1t
- Paper: https://arxiv.org/abs/2506.05209

Collaborators

poolside

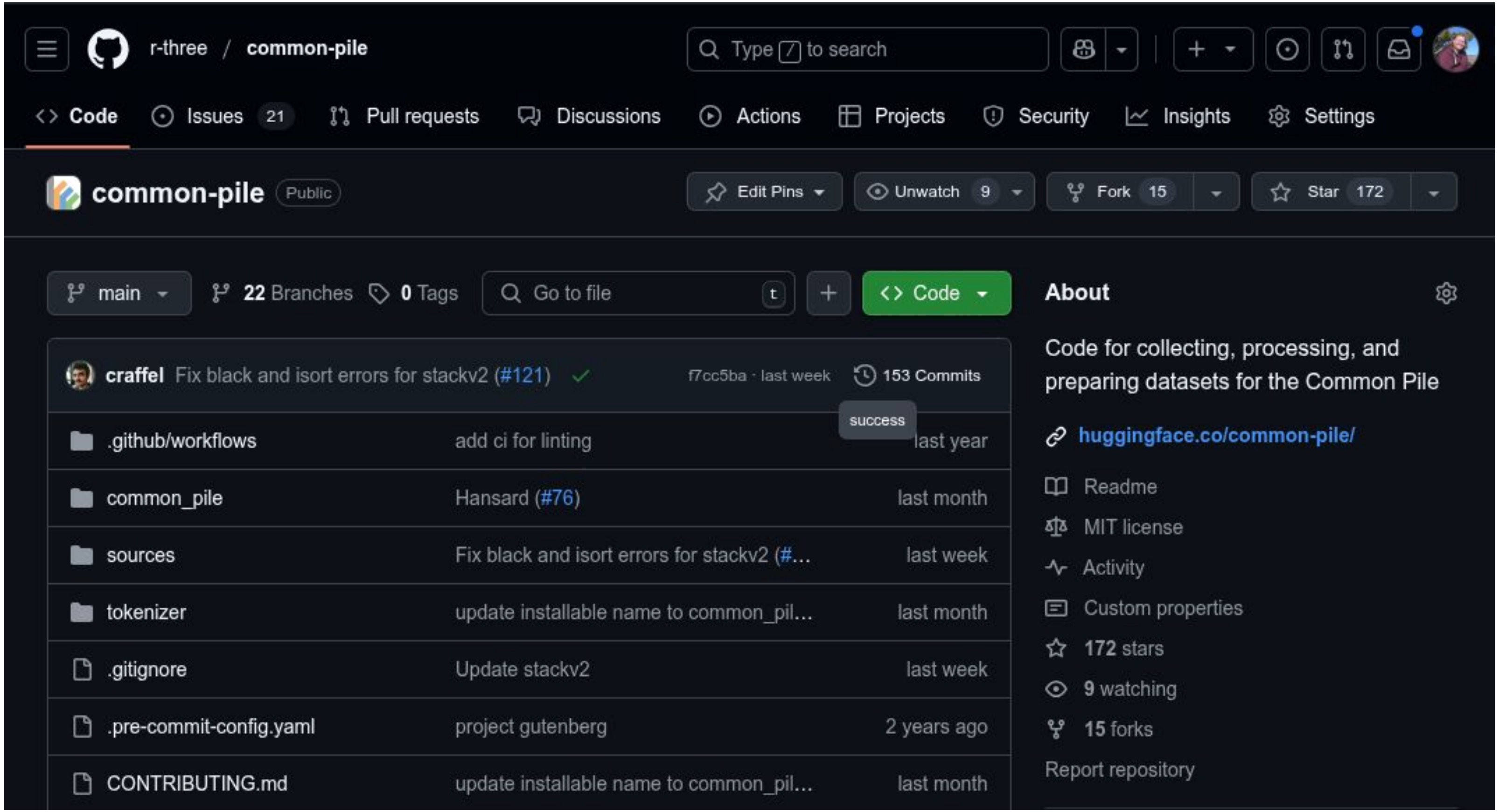
EleutherAI

Ai2

VECTOR INSTITUTE

You?

Contribute—https://github.com/r-three/common-pile



craffel Fix black and isort errors for stackv2 (#121) ✓ f7cc5ba · last week 153 Commits

.github/workflows	add ci for linting	last year
common_pile	Hansard (#76)	last month
sources	Fix black and isort errors for stackv2 (#...	last week
tokenizer	update installable name to common_pil...	last month
.gitignore	Update stackv2	last week
.pre-commit-config.yaml	project gutenberg	2 years ago
CONTRIBUTING.md	update installable name to common_pil...	last month

About

Code for collecting, processing, and preparing datasets for the Common Pile

huggingface.co/common-pile/

- Readme
- MIT license
- Activity
- Custom properties
- 172 stars
- 9 watching
- 15 forks

Report repository