The Data Provenance Initiative

A Large Scale Audit of Dataset Licensing & Attribution in AI

Shayne Longpre, Naana Obeng-Marnu, William Brannon

Authors: Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, Will Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Alexis Wu, Enrico Shippole, Kurt Bollacker, Sherry Wu, Luis Villa, Sandy Pentland, Sara Hooker

Shayne Longpre
MIT Media Lab
Human Dynamics
slongpre@media.mit.edu

William Brannon
MIT Center for Constructive
Communication, Media Lab
wbrannon@media.mit.edu

Naana Obeng-Marnu
MIT Center for Constructive
Communication, Media Lab
naanaom@media.mit.edu

Agenda

1) **What & Why Data Provenance?**
2) **What did we collect?**
3) **The Explorer**
4) **Ecosystem Audit & Analysis**
    - Commercial vs Non-Commercial divide
    - Global North vs Global South
    - Legal situation going forward
5) **What's Next?**

What & Why Data Provenance?

(Motivation)

What data *should* we use for training?

1) What is right for our application? (tasks, topics, domains, languages)
2) What is legally permissible? (sources, licenses, terms, precedence of use)
3) What satisfies ethical/PR concerns? (creators, representation)

What data *should* we use for training?

Terms of Use

OpenAI suspends ByteDance's account after it allegedly used GPT to build rival AI product: report

- The NYPost, Dec 18

Non-consensual graphic data

TECH / ARTIFICIAL INTELLIGENCE

AI image training dataset found to include child sexual abuse imagery

- The Verge, Dec 20

Copyright

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

- New York Times, Dec 27

What data *should* we use for training?

Synthetic
Terms of Use
Privacy
Languages
Time of Collection
Accuracy
Creators
Non-consensual graphic data
Topics
Domains
Tasks
Sources
Size metrics
Copyright
Popularity

OpenAI suspends ByteDance's account after it allegedly used GPT to build rival AI product: report

- The NYPost, Dec 18

TECH / ARTIFICIAL INTELLIGENCE

AI image training dataset found to include child sexual abuse imagery

- The Verge, Dec 20

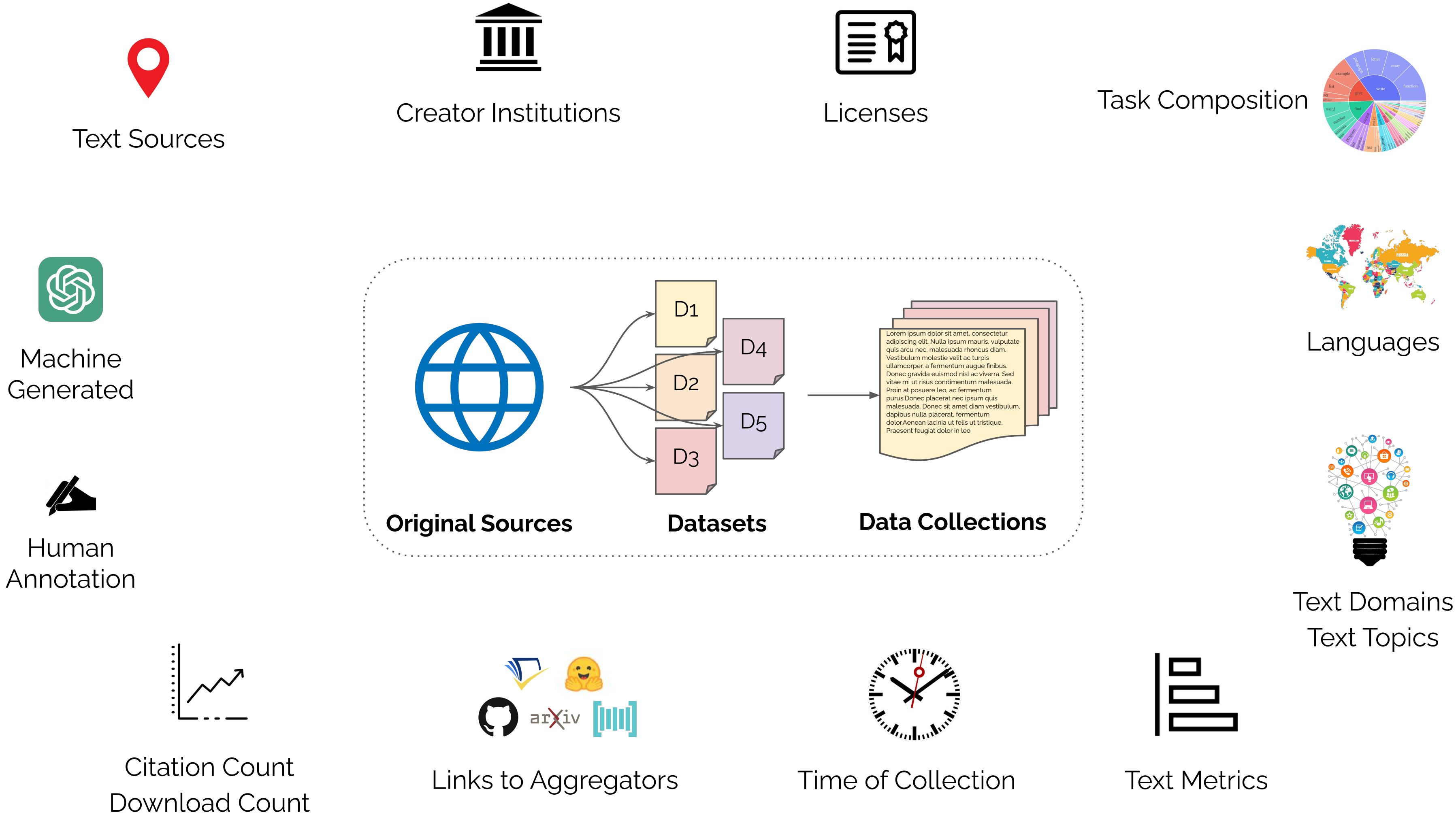
The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

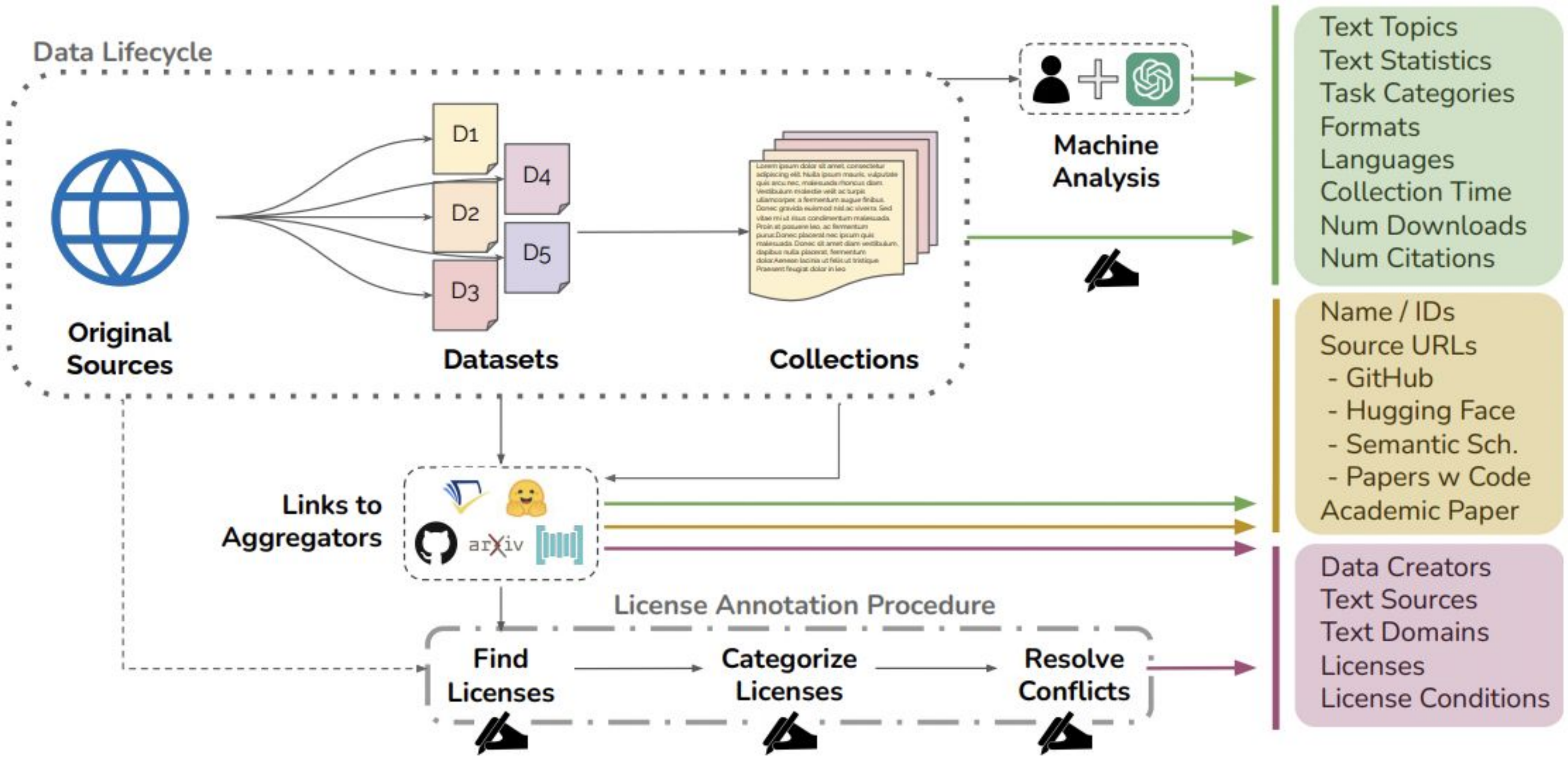
- New York Times, Dec 27

65%

Of the datasets have incorrect licenses

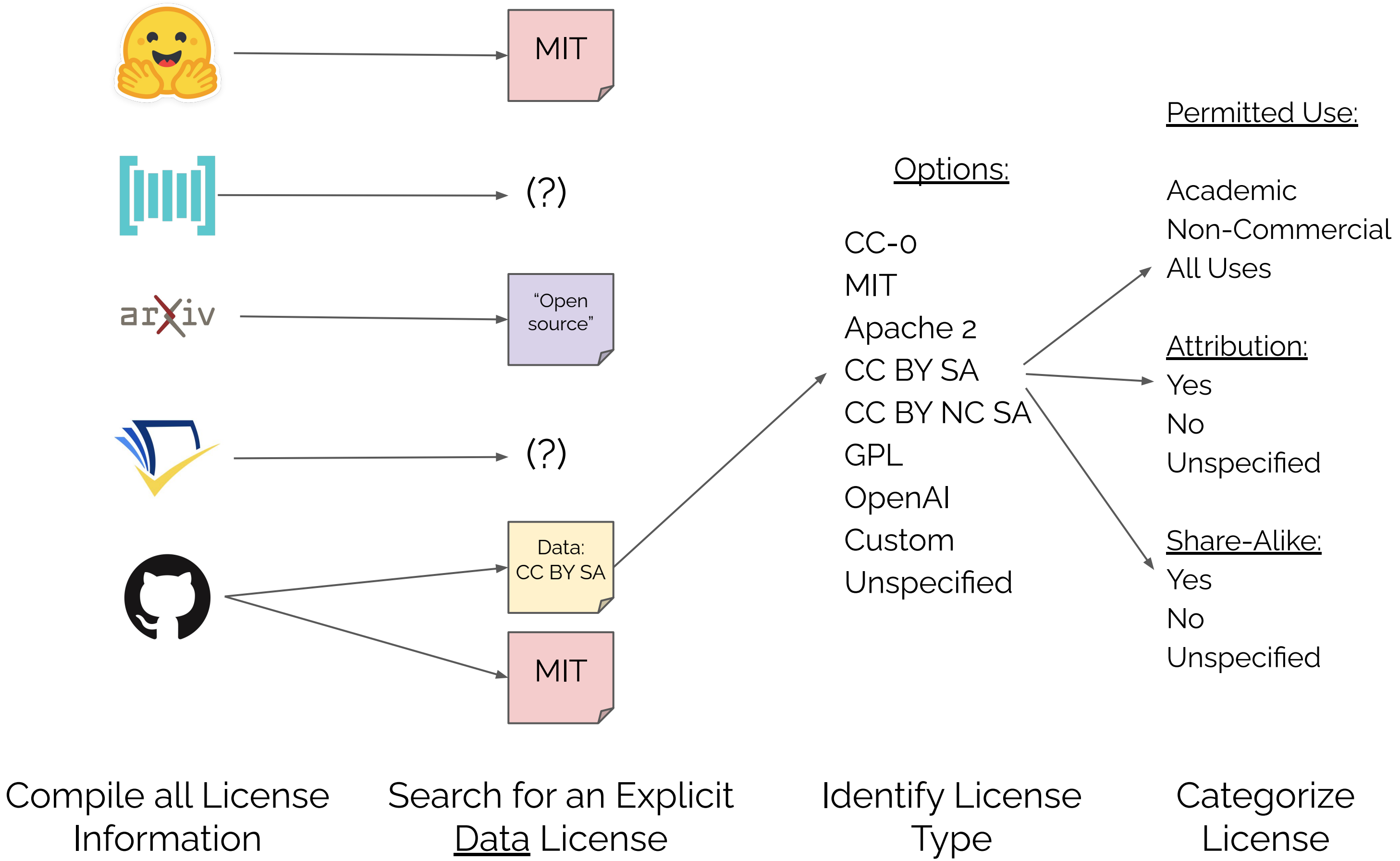
We face a crisis in data


Text Sources
Creator Institutions
Licenses
Task Composition
Machine Generated
Human Annotation
Languages
D1
D2
D3
D4
D5
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla ipsum mauris, vulputate quis arcu nec, malesuada rhoncus diam. Vestibulum molestie velit ac turpis ullamcorper, a fermentum augue finibus. Donec gravida euismod nisl ac viverra. Sed vitae mi ut risus condimentum malesuada. Proin at posuere leo, ac fermentum purus.Donec placerat nec ipsum quis malesuada. Donec sit amet diam vestibulum, dapibus nulla placerat, fermentum dolor.Aenean lacinia ut felis ut tristique. Praesent feugiat dolor in leo
Original Sources
Datasets
Data Collections
Text Domains
Text Topics
Citation Count
Download Count
arXiv
Links to Aggregators
Time of Collection
Text Metrics



Data Lifecycle
Original Sources
D1
D2
D3
D4
D5
Datasets
Collections
Machine Analysis
Text Topics
Text Statistics
Task Categories
Formats
Languages
Collection Time
Num Downloads
Num Citations
Name / IDs
Source URLs
- GitHub
- Hugging Face
- Semantic Sch.
- Papers w Code
Academic Paper
Links to Aggregators
arXiv
Data Creators
Text Sources
Text Domains
Licenses
License Conditions
License Annotation Procedure
Find Licenses
Categorize Licenses
Resolve Conflicts

License Annotation Workflow



A crisis in data…

Provenance

- Missing/ambiguous licenses 65% of HF licenses are missing or incorrect.
- License revisions post-release (e.g. Mosaic ML's MPT models)
- Lawsuits (e.g. Stability AI and OpenAI)

Transparency

- Undocumented data
- Inability to properly audit
- E.g. Abilities, copyright, originality, PII leaks, train/test overlap…

Understanding

- Biases & Toxicity
- Unintended model behavior
  - Chatty or terse?
  - Cautious or uninhibited?
  - Mono- or Multilingual?
- => Poor quality models

Where does our work fit in?

Data Nutrition Project

- Chmielinski et al 2020
- Standardized labels for many kinds of datasets

Datasheets for Datasets

- Gebru et al 2018
- Standardized labels for ML data

Ecosystem Graphs

- Bommasani et al 2023
- Mapping model / data / application relations

We go beyond these by...

Actually annotating + preparing data!

What did we collect?

(and what did we build)

A crisis in data

provenance	***transparency***	***understanding***

A Large-Scale Dataset Audit
(1858+ Datasets)

Trace provenance lineage for each dataset, from text sources to dataset creators, to licenses.

A **data explorer tool** for developers to filter on any data provenance or characteristics criteria, download, and generate a Data Provenance Card for attribution.

The **largest empirical analysis** of supervised text data, and their provenance, to date.

Collection	Property Counts							Text Lens		Dataset Types							
	Datasets	Dialogs	Tasks	Langs	Topics	Domains	Downs	Inpt	Tgt	Source	Z	F	C	R	M	Use	O
Airoboros	1	17k	5	2	10	1	1k	347	1k	🤖	✔					🔴	✔
Alpaca	1	52k	8	1	10	1	100k	505	270	🤖	✔					🔴	✔
Anthropic HH	1	161k	3	1	10	1	82k	69	311	🤖				✔		🔵	
BaizeChat	4	210k	12	2	37	3	<1k	74	234	🤖	✔					🔴	✔
BookSum	1	7k	4	1	10	1	<1k	14k	2k	🌐	✔					🔴	
CamelAI Sci.	3	60k	2	1	29	1	<1k	190	2k	🤖	✔					🔴	✔
CoT Coll.	6	2,183k	12	7	29	1	<1k	728	265	🤖			✔			🔴	✔
Code Alpaca	1	20k	3	2	10	1	5k	97	196	🤖	✔					🟡	✔
CommitPackFT	277	702k	1	278	751	1	4k	645	784	🌐	✔					🔵🟡	
Dolly 15k	7	15k	5	1	38	1	10,116k	423	357	🌐	✔					🔵	
Evol-Instr.	2	213k	11	2	17	1	2k	570	2k	🤖	✔					🔴	✔
Flan Collection	450	9,813k	19	39	1k	23	19k	2k	128	🌐🤖	✔	✔	✔			🔵🟡🔴	✔
GPT-4-Alpaca	1	55k	7	1	10	1	1k	130	543	🤖	✔					🔴	✔
GPT4AllJ	7	809k	10	1	56	1	<1k	883	1k	🤖	✔					🟡🔴	✔
GPTeacher	4	103k	8	2	33	1	<1k	227	360	🤖	✔					🔵	✔
Gorilla	1	15k	4	2	10	2	<1k	119	76	🤖	✔					🔵	✔
HC3	12	37k	6	2	102	6	2k	119	652	🤖				✔		🔵🟡🔴	✔
Joke Expl.	1	<1k	2	1	10	1	<1k	96	547	🌐	✔					🔵	
LAION OIG	26	9,211k	12	1	171	11	<1k	343	595	🌐🤖					✔	🔵🟡	✔
LIMA	5	1k	10	2	43	6	3k	228	3k	🌐	✔	✔			✔	🔴	
Longform	7	23k	11	1	63	4	3k	810	2k	🤖	✔					🔵🟡	✔
OpAsst OctoPack	1	10k	3	20	10	1	<1k	118	884	🌐					✔	🔵	
OpenAI Summ.	1	93k	5	1	10	1	14k	1k	134	🤖				✔		🔵	✔
OpenAssistant	19	10k	4	20	99	1	14k	118	711	🌐					✔	🔵	
OpenOrca	4	4,234k	11	1	30	23	28k	1k	492	🤖	✔					🔵 🔴	✔
SHP	18	349k	6	2	151	1	4k	824	496	🌐				✔		🟡	
Self-Instruct	1	83k	6	2	10	1	3k	134	104	🤖	✔					🔵	✔
ShareGPT	1	77k	9	1	10	2	<1k	303	1k	🤖					✔	🟡	✔
StackExchange	1	10,607k	1	2	10	1	<1k	1k	901	🌐	✔					🟡	
StarCoder	1	<1k	1	2	10	1	<1k	195	504	🤖	✔					🔵	
Tasksource Ins.	288	3,397k	13	1	582	20	<1k	518	18	🌐🤖	✔					🔵🟡🔴	✔
Tasksource ST	229	338k	15	1	477	18	<1k	3k	6	🌐🤖	✔					🔵🟡🔴	✔
TinyStories	1	14k	4	1	10	1	12k	517	194k	🤖	✔					🔵	✔
Tool-Llama	1	37k	2	2	10	1	-	7k	1k	🤖					✔	🔴	✔
UltraChat	1	1,468k	7	1	11	2	2k	282	1k	🤖	✔				✔	🔴	✔
Unnatural Instr.	1	66k	4	1	10	1	<1k	331	68	🤖	✔					🔵	✔
WebGPT	5	20k	4	1	35	3	1k	737	743	🤖				✔		🔵	✔
xP3x	467	886,240k	5	245	151	14	<1k	589	441	🌐🤖	✔					🔵🟡🔴	

Anthropic HH-RLHF

Dolly v2

Baize Data

Lima

Flacuna

The Flan Collection

Tool LLaMA

Evol-Instruct v2

ShareGPT

WebGPT
OpenAI Summarization

Super-Natural Instructions

Camel Science

Orca

UltraChat

And More…

Starcoder Self-Instruct
Unnatural Instructions
Joke Explanations
Book Summaries
CoT collection
Self-Instruct
GPTeacher
Longform
GPT-4All
Airboros
SHP …

Open Assistant
Open Instruction Generalist

Alpaca
CodeAlpaca

Tasksource

xP3

The Explorer

www.dataprovenance.org

- Open source: who is this tool for?
- Gif: maybe not dark mode. Select many datasets so graphs looks maximally interesting
- What is the data sources? Who are the creators?
- **Heatmaps: languages → creators**

The Data Provenance Explorer *[dataprovenance.org]*

- ***Model Trainers can filter for datasets that match their use case and confirm its licensing to source their datasets legally and ethically***

- ***Researchers can search for patterns and uncover biases***

- ***Data Creators can verify how their data is being used***

Data Provenance Explorer

The Data Provenance Initiative is a large-scale audit of AI datasets used to train large language models. As a first step, we've traced 1800+ popular, text-to-text finetuning datasets from origin to creation, cataloging their data sources, licenses, creators, and other metadata, for researchers to explore using this tool. The purpose of this work is to improve transparency, documentation, and informed use of datasets in AI.

You can download this data (with filters) directly from the Data Provenance Collection.

If you wish to contribute or discuss, please feel free to contact the organizers at data.provenance.init@gmail.com.

NB: It is important to note we collect *self-reported licenses*, from the papers and repositories that released these datasets, and categorize them according to our best efforts, as a volunteer research and transparency initiative. The information provided by any of our works and any outputs of the Data Provenance Initiative do **NOT**, and are **NOT** intended to, constitute legal advice; instead, all information, content, and materials are for general informational purposes only.

Data Repository

Paper

DATA PROVENANCE INITIATIVE

Instructions

Expand for Instructions!

Select the preferred criteria for your datasets.

Select the datasets licensed for these use cases

Academic-Only

Commercial

Academic-Only

Include Datasets w/ Attribution Requirements

Include Datasets w/ Share Alike Requirements

Always include datasets w/ OpenAI-generated data. (I.e. See `instructions` above for details.)

Select the languages to cover in your datasets

All ×

Select data release time constraints

2000-01-01

2023-12-01

1999-12-18

2023-12-01

Select the task categories to cover in your datasets

All ×

Select the domain types to cover in your datasets

All ×

Submit Selection

Data Summary | Global Representation 🌍 | Text Characteristics ✏️ | Data Licenses | Inspect Individual Datasets 🔍

General Properties of your collection

Given your selection, see the quantity of data (collections, datasets, dialogs), the characteristics of the data (languages, tasks, topics), and the sources of data covered (sources, domains, % synthetically generated by models).

40
/ 44
Collections

524
/ 524
Languages

23
/ 23
Text Domains

1786
/ 1858
Datasets

225
/ 253
Task Categories

388
/ 430
Text Sources

920739953
/ 930750141
Dialogs

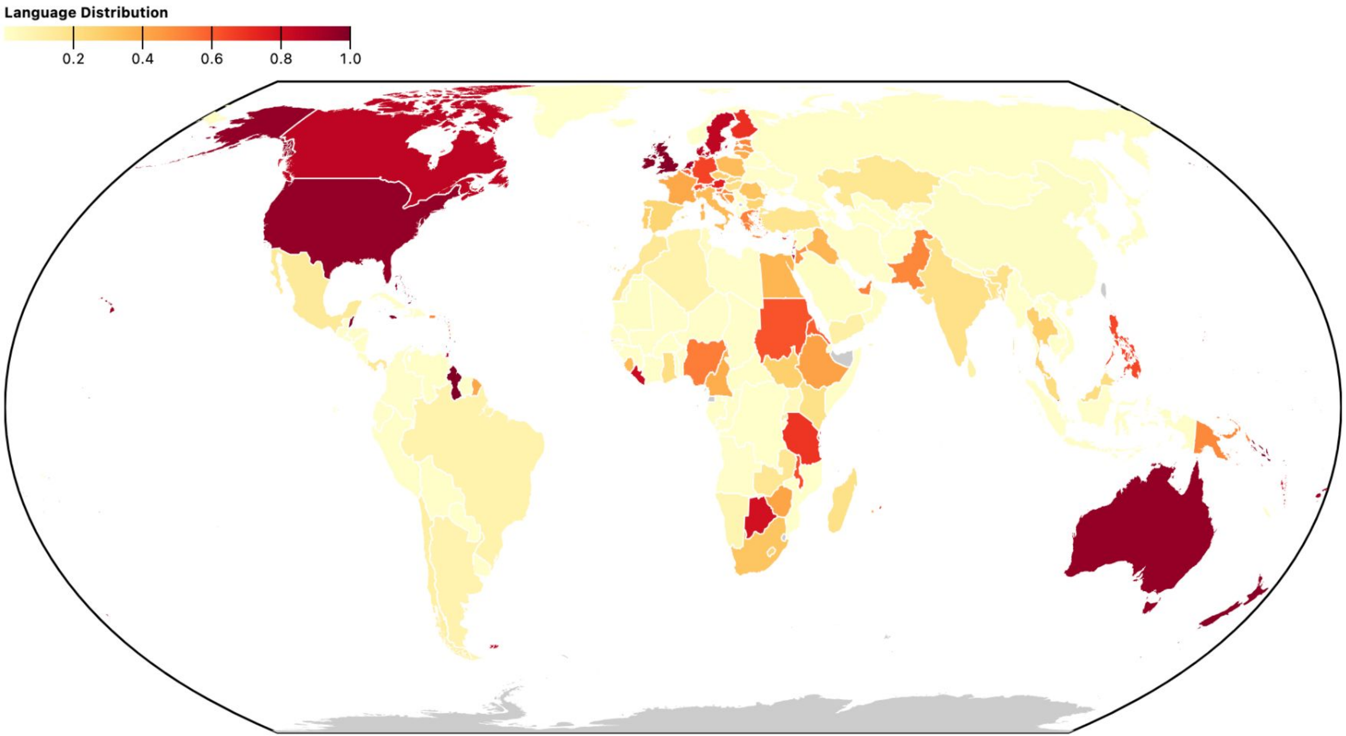
2471
/ 2532
Topics

14.0
% Synthetic Text

Summary of Data Collections

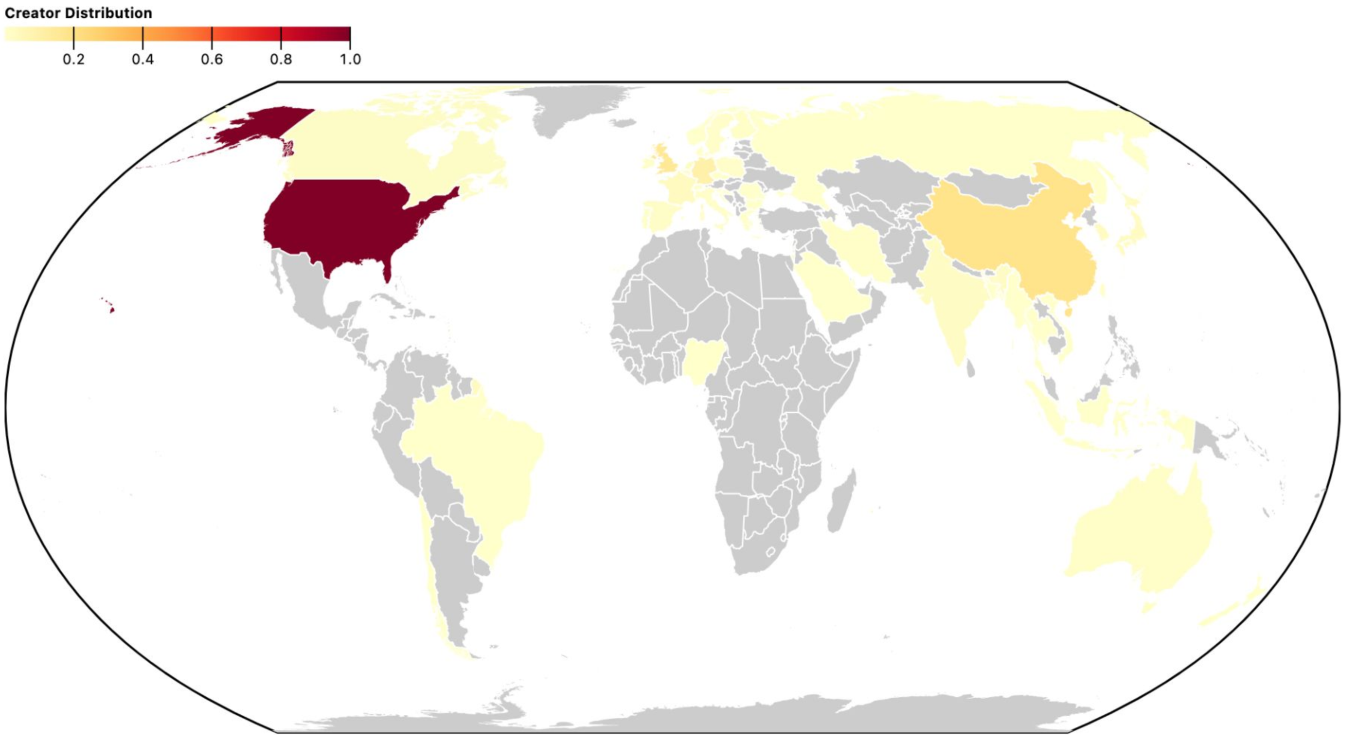
Collection	# Datasets	# Exs	# Languages	# Tasks	# Topics	# Sources	Generated By	Mean Input Chars	Mean Target Char
CoT Collection	6	2182808	7	14	29	1	OpenAI Codex	459	192.
Stanford Human Preferences	18	348718	2	6	151	1		925	562.

Language Representation by Country



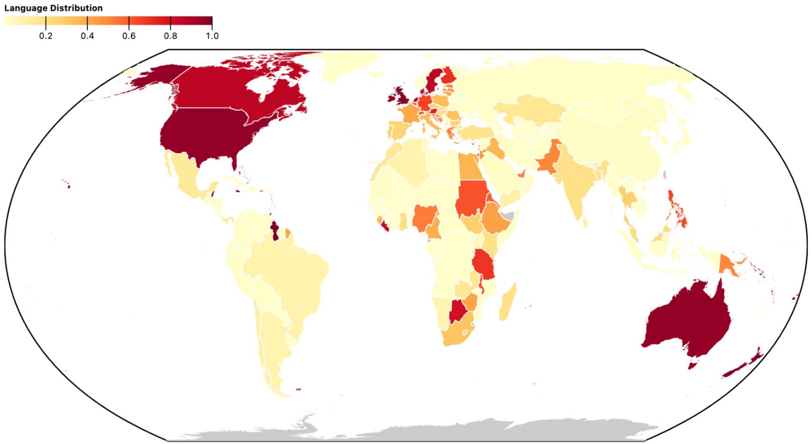
* From dataprovenance.org with filters [Non-Commercial Licenses, Indo-European Languages]

Dataset Creator Representation by Country

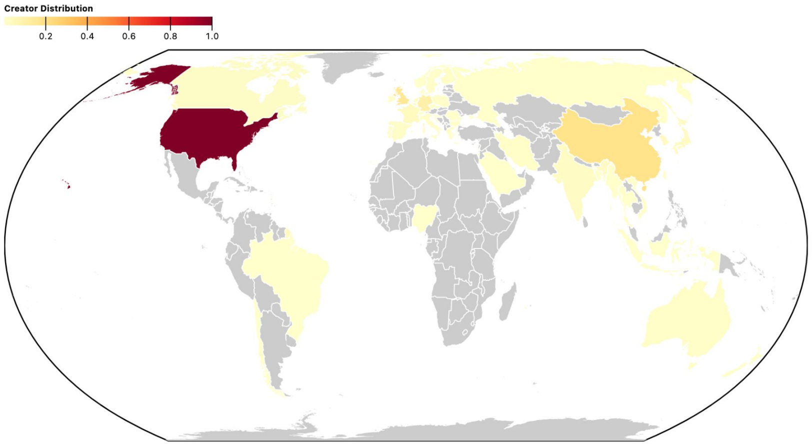


* From dataprovenance.org with filters [Non-Commercial Licenses, Indo-European Languages]

Languages vs Creators Side by Side



Language Representation by Country

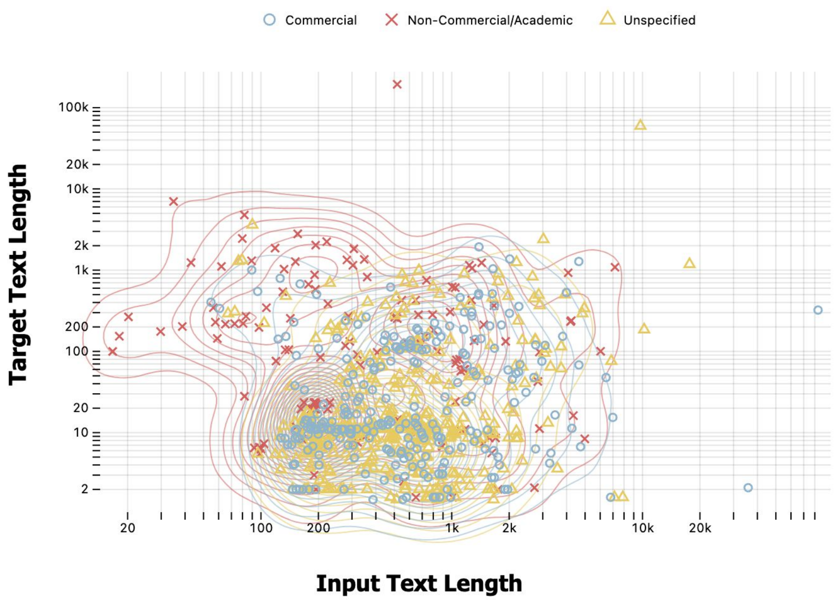


Dataset Creator Representation by Country

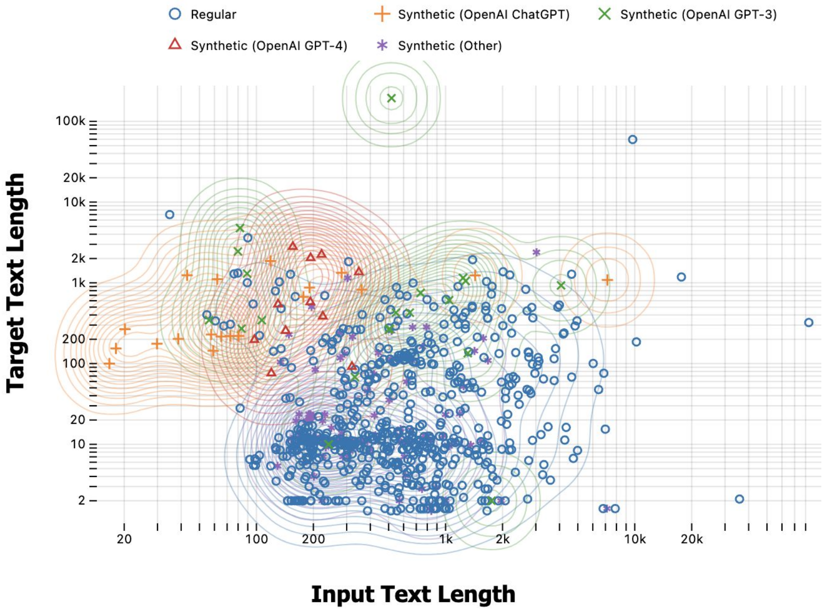
* From dataprovenance.org with filters [Non-Commercial Licenses, Indo-European Languages]

A Global ***minority*** is driving the creation of datasets that affect us ***all***.

Text Length Metrics



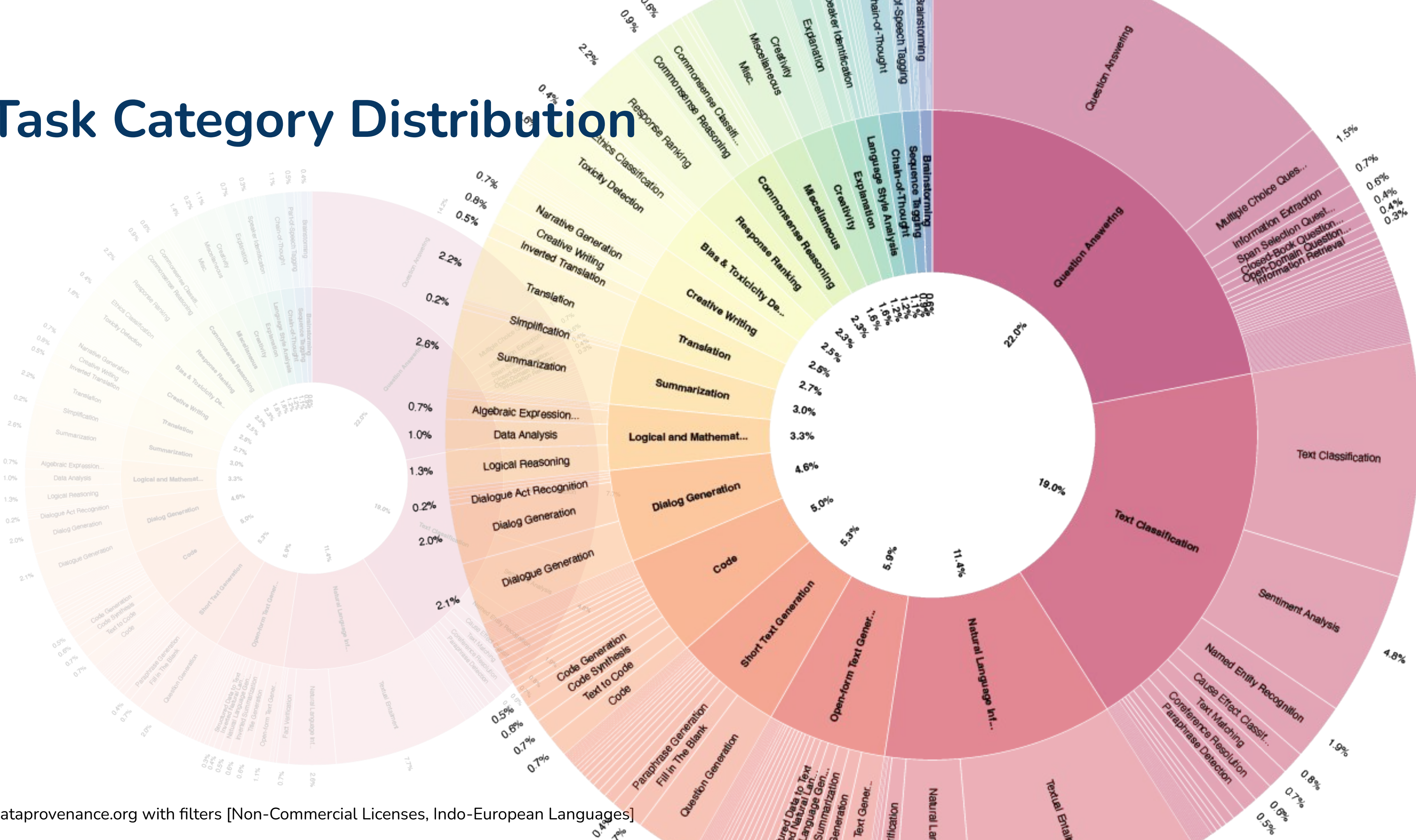
By License



By Regular/Synthetic Text

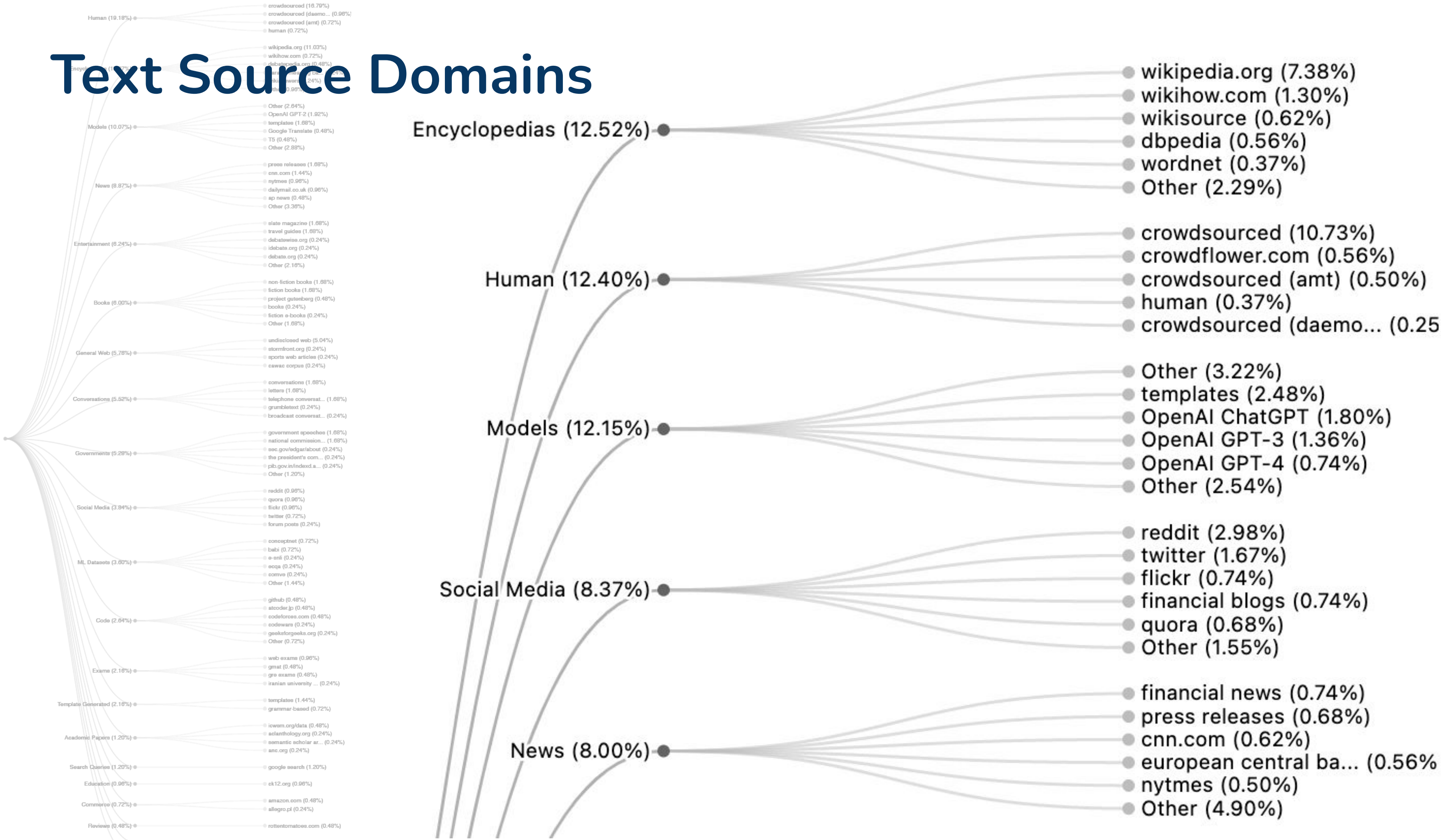
* From dataprovenance.org with filters [Non-Commercial Licenses, Indo-European Languages]

Task Category Distribution



* From dataprovenance.org with filters [Non-Commercial Licenses, Indo-European Languages]

Text Source Domains



* From dataprovenance.org with filters [Non-Commercial Licenses, Indo-European Languages]

Try it yourself!

Ecosystem Analysis & Takeaways

(Why does it matter?)

RQ1: How accurate is public license info?

43.4%

53.9%

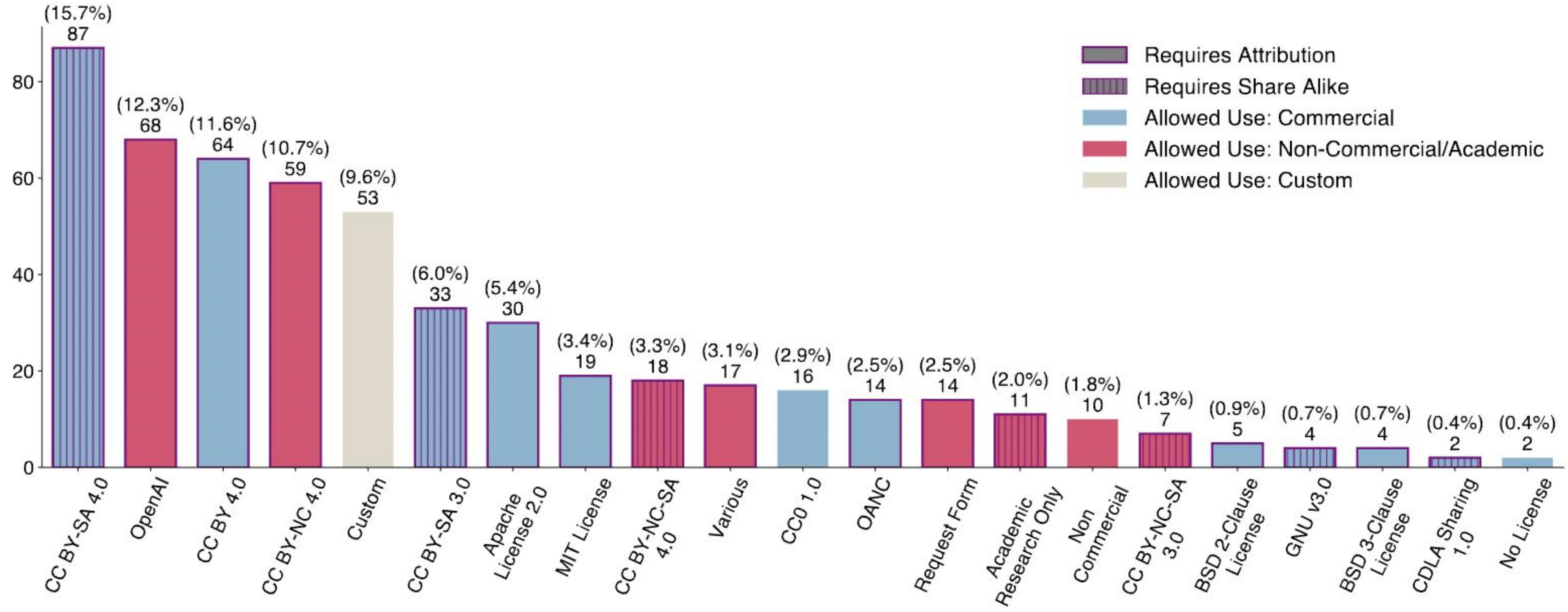
35.3%

(% Correct Category)

Unspecified: 69%-72%
After annotation: **30.7%**

Not very accurate!

RQ2: What’s available by license type?

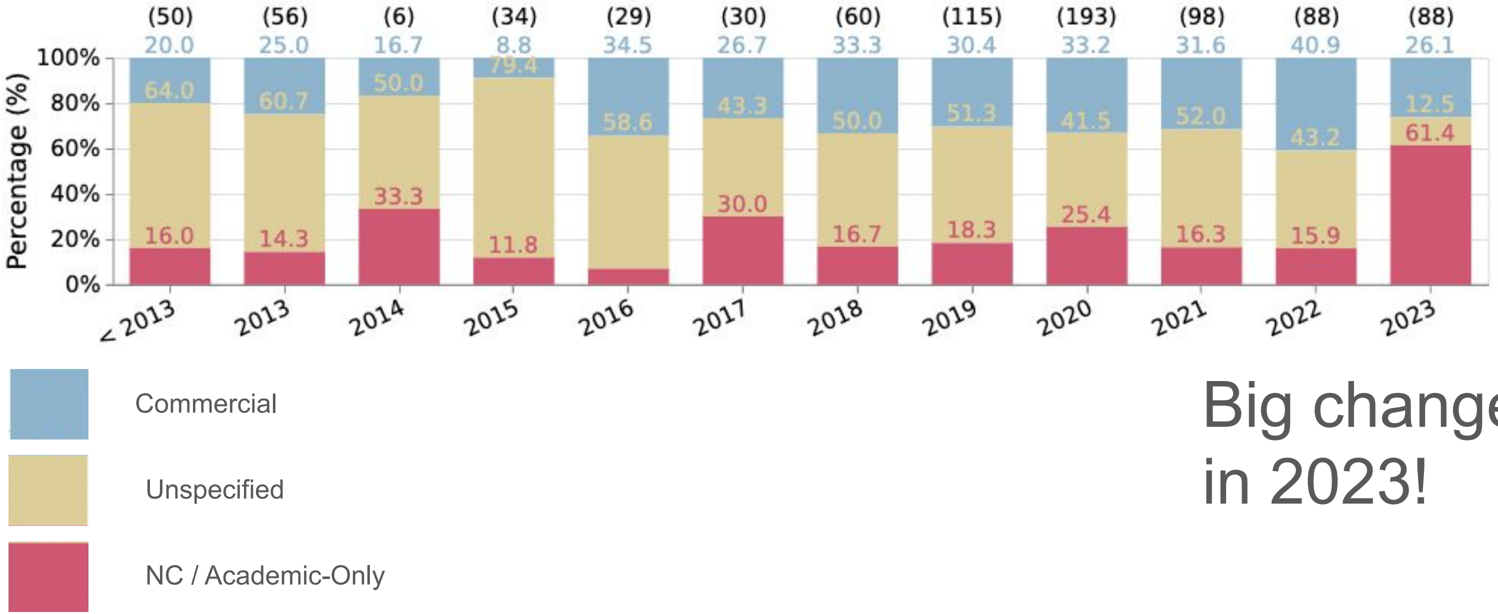


Mostly commercially permissive!

SA: 33% **BY: 73%**

Long tail of custom licenses

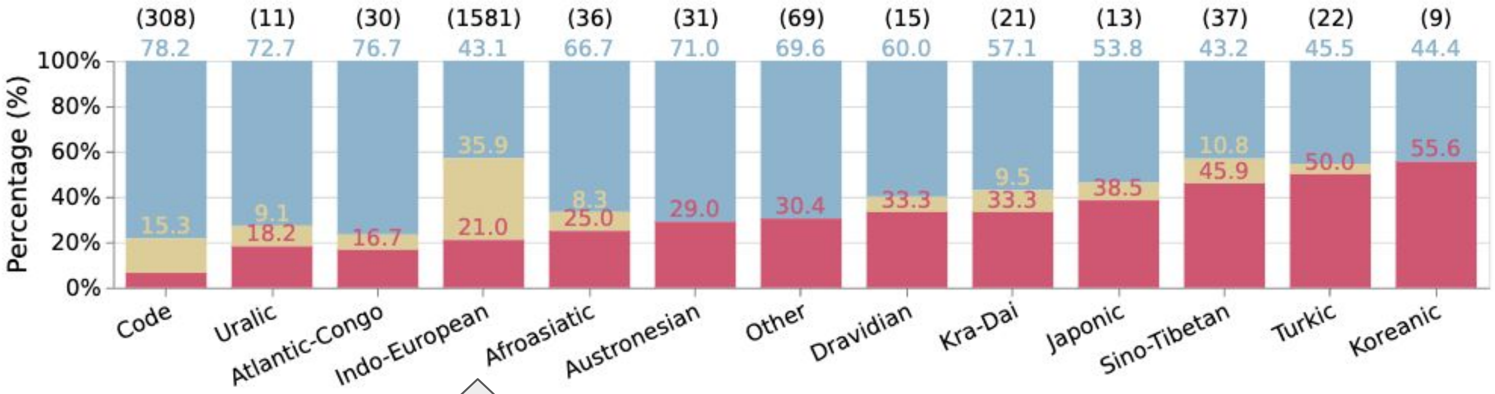
RQ3: Changes over time?



Big changes in 2023!

New datasets are increasingly noncommercial

RQ4(a): Disparities by language?



Commercial

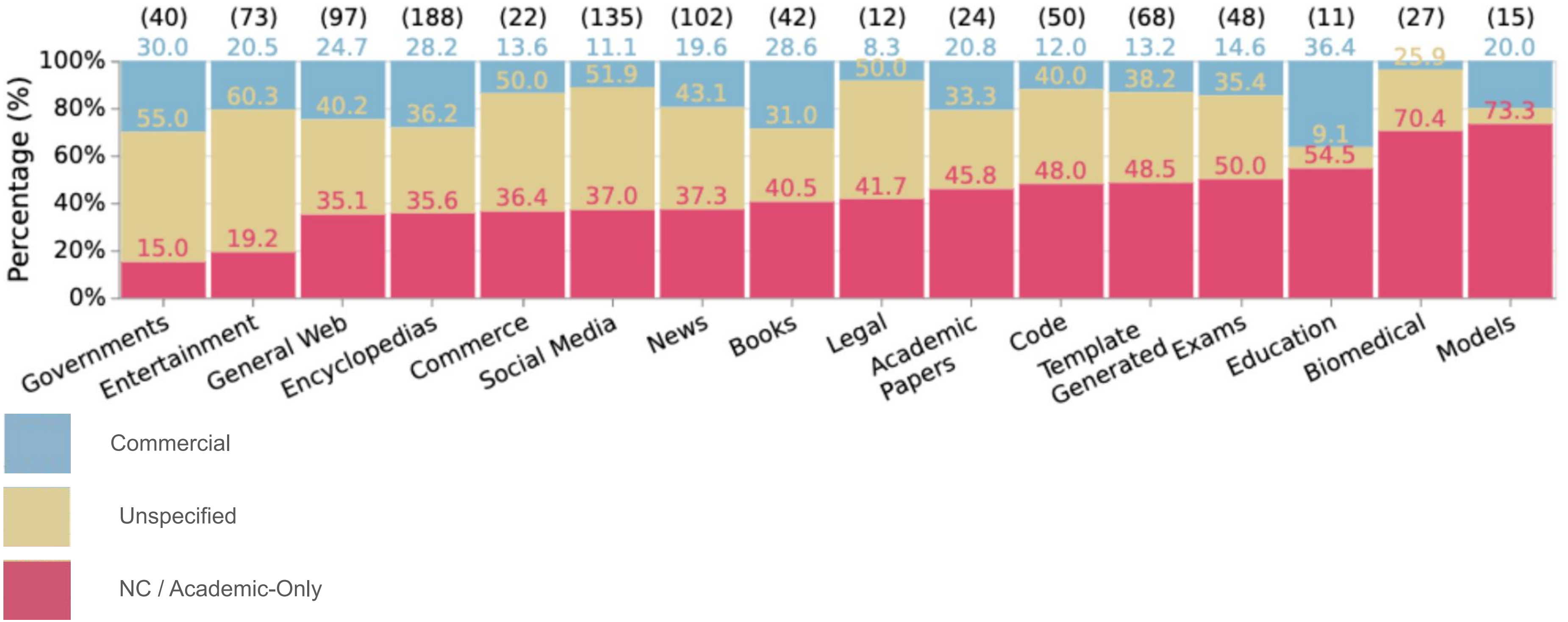
Unspecified

NC / Academic-Only

“Unspecified” especially for Indo-European languages

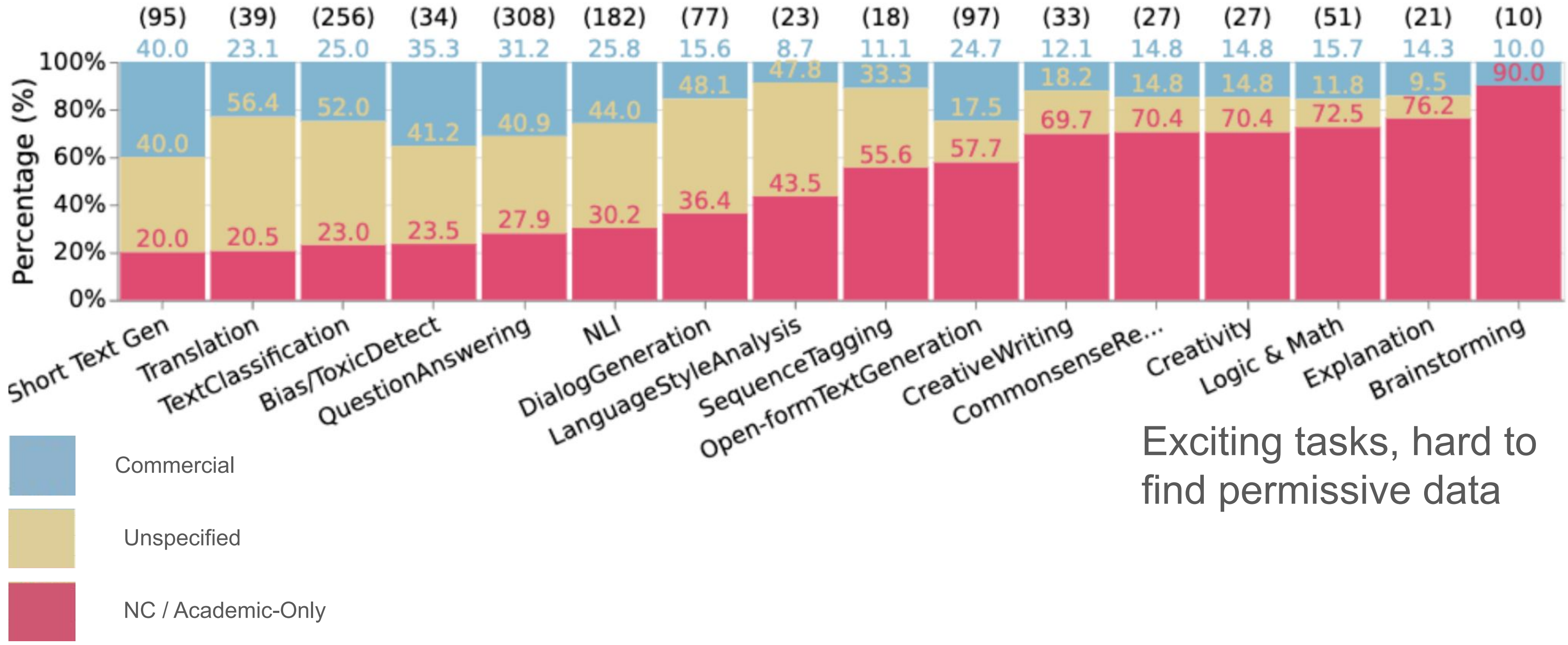
Less commercial data for low-resource languages!

RQ4(b): License differences by domain?



Big differences between domains!

RQ4(c): License differences by task?



Exciting tasks, hard to find permissive data

Even bigger inter-task differences!

Correct license		License according to aggregators (Agg.)				
License	Count	Agg.	Comm.	Unspec.	Non-Comm.	Acad.-Only
Commercial	856 (46.1%)		349	507	0	0
			176	677	1	2
			313	520	1	22
Unspecified	570 (30.7%)		112	458	0	0
			164	395	6	5
			31	523	1	15
Non-Commercial	352 (19.0%)		49	303	0	0
			113	152	80	7
			2	191	157	2
Academic-Only	80 (4.3%)		9	71	0	0
			9	65	2	4
			5	65	2	8
Total	1858 (100%)		519 (28%)	1339 (72%)	0 (0%)	0 (0%)
			462 (25%)	1289 (69%)	89 (5%)	18 (1%)
			351 (19%)	1299 (70%)	161 (9%)	47 (3%)

Table 2: **The distribution of license use categories shows our licenses have far fewer "Unspecified" omissions than GitHub (, 72%), Hugging Face (, 69%), and Papers with Code (, 70%), categorizing license more confidently into commercial or non-commercial categories.** GitHub, Hugging Face, and Papers with Code match our licenses (green regions) 43%, 35%, and 54% of the time, respectively, and suggest incorrect licenses that are *too permissive* 29%, 27%, and 16% of the time.

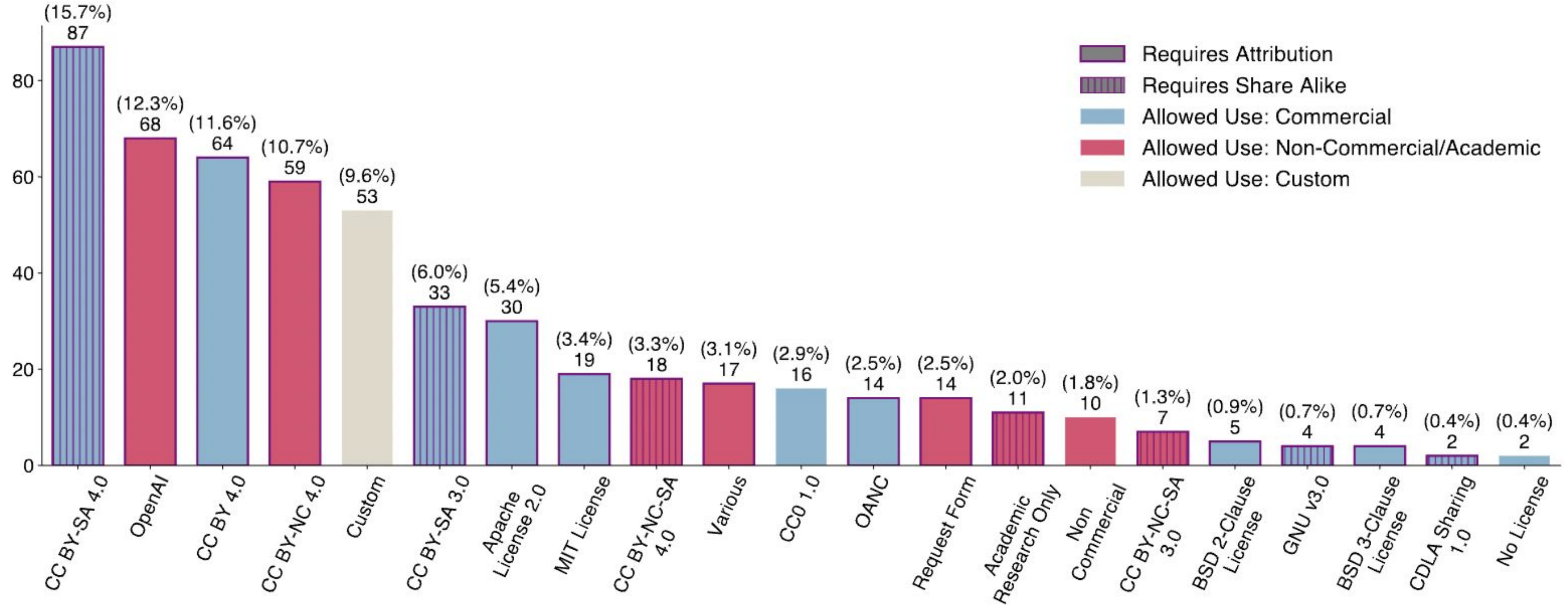


Figure 2: **We plot the distributions of licenses used in the DPCollection, a popular sample of the major supervised NLP datasets.** We find a long tail of custom licenses, adopted from software for data. 73% of all licenses require attribution, and 33% share-alike, but the most popular are usually commercially permissive.

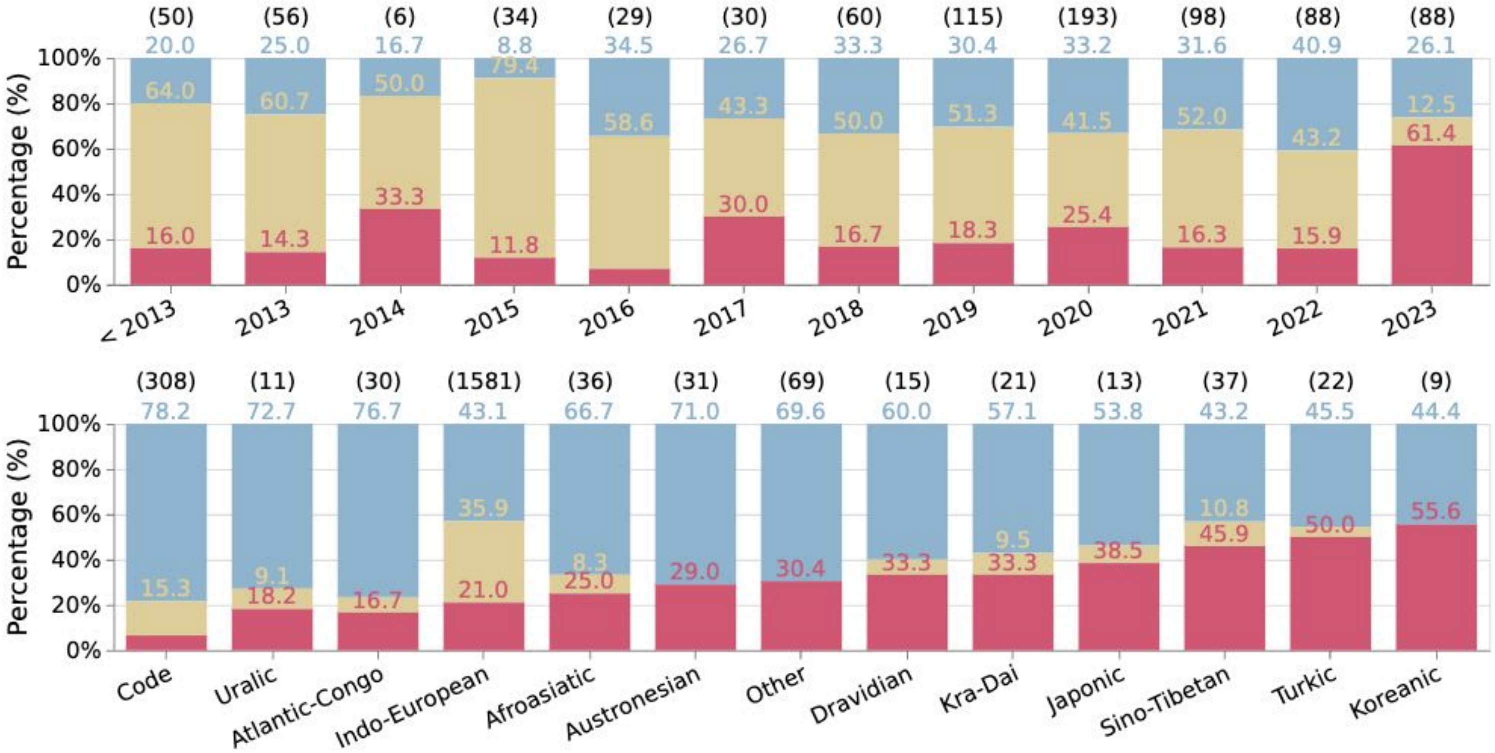


Figure 3: The distribution of datasets in each time of collection (top) and language family (bottom) category, with total count above the bars, and the portion in each license use category shown via bar color. Red is Non-commerical/Academic-Only, Yellow is Unspecified, and Blue is Commercial. **Lower resource languages, and datasets created in 2023 see a spike in non-commercial licensing.**

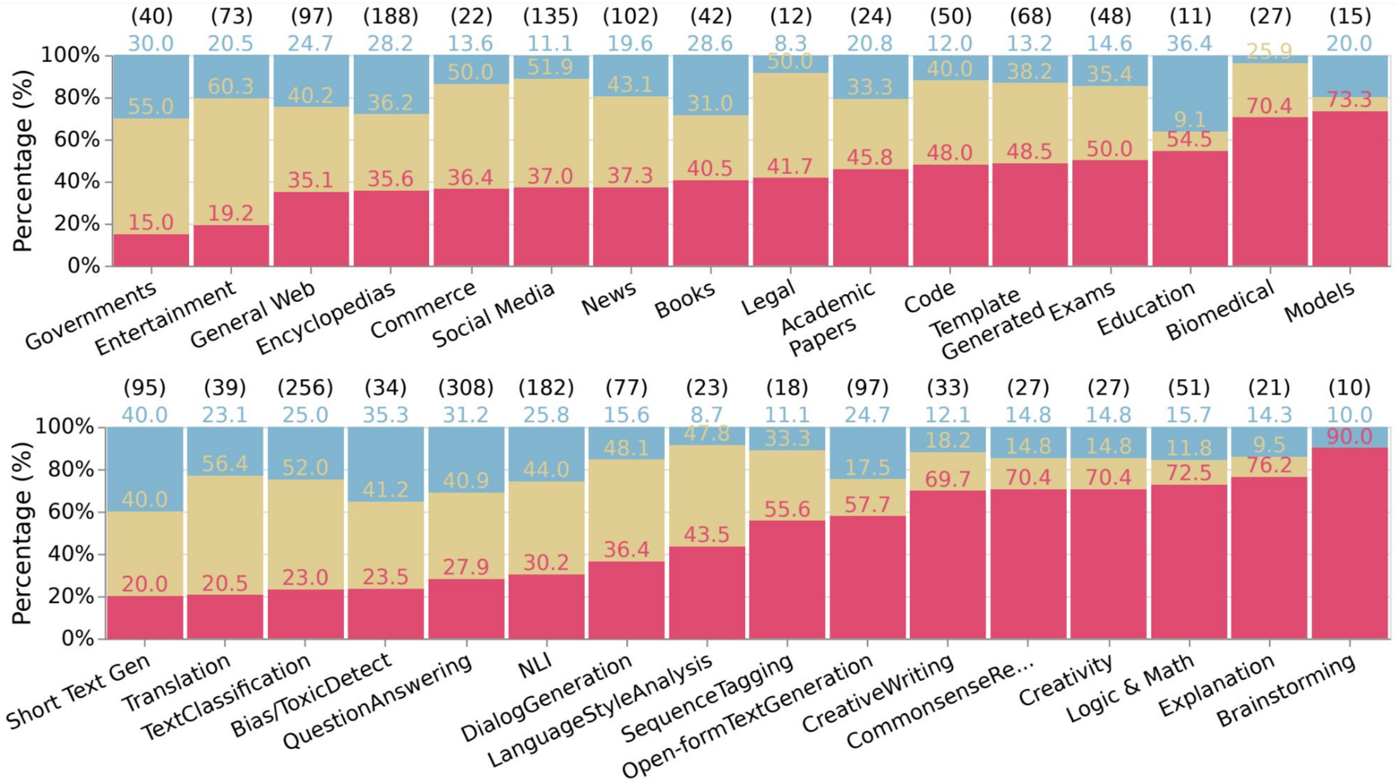


Figure 4: The distribution of datasets in each **Domain Source (top)** and **task (bottom)** category, with total count above the bars, and the portion in each license use category shown via bar color. Red is Non-commerical/Academic-Only, Yellow is Unspecified, and Blue is Commercial. **Creative, reasoning, and long-form generation tasks, as well as datasets sourced from models, exams, and the general web see the highest rate of non-commercial licensing.**

What's Next?

We are growing...

- 40-50+ contributors
- 15+ countries & organizations
- + Vision + Speech + Pre-training Datasets
- Wider ecosystem audit
- Investigating data accessibility, representation & composition

AI researchers uncover ethical, legal risks to using popular data sets

The Data Provenance Initiative analyzed data sets used to build generative AI and found confusion surrounding licensing and fair use

- Washington Post, Oct 25

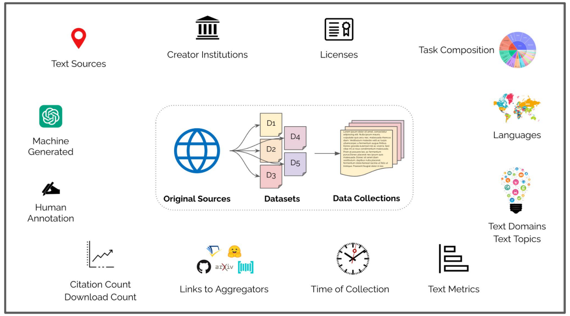
MIT, Cohere for AI, others launch platform to track and filter audited AI datasets

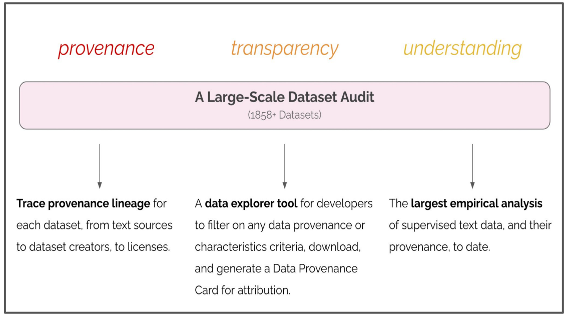
- VentureBeat, Oct 25

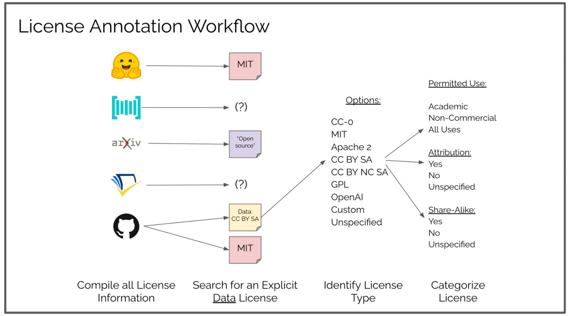
Public AI Training Datasets Are Rife With Licensing Errors › An audit of popular datasets suggests developers face legal and ethical risks

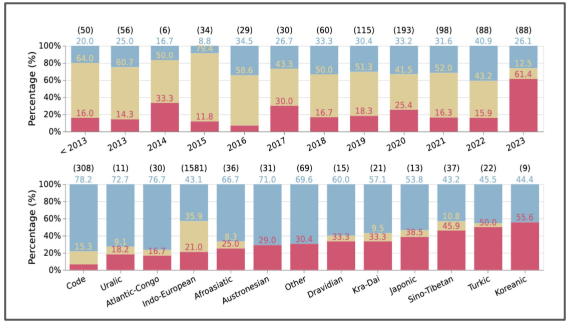
- IEEE Spectrum, Nov 8

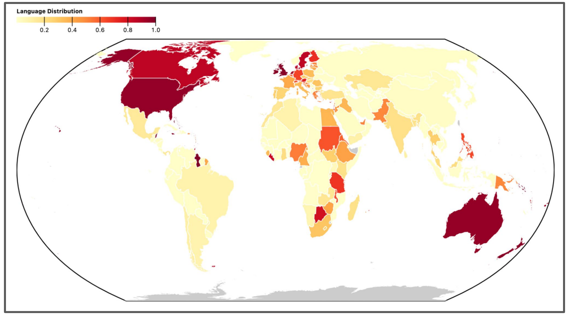
dataprovenance.org

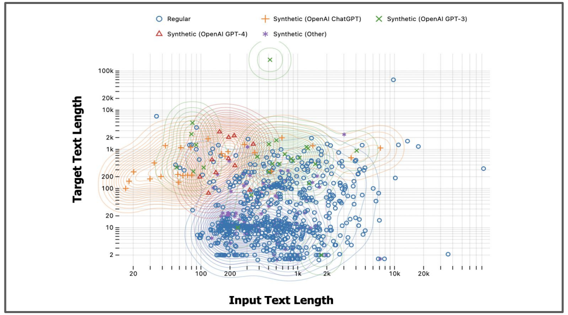












Thank you!

dataprovenance.org