A large-scale audit of dataset licensing and attribution in AI

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al.

nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00878-8

A large-scale audit of dataset licensing and attribution in AI

Received: 15 February 2024

Accepted: 10 July 2024

Published online: 30 August 2024

Check for updates

Shayne Longpre[1,15], Robert Mahari[1,2,15] ✉, Anthony Chen[3], Naana Obeng-Marnu[1,4], Damien Sileo[5], William Brannon[1,4], Niklas Muennighoff[6], Nathan Khazam[7], Jad Kabbara[1,4], Kartik Perisetla[8], Xinyi (Alexis) Wu[9], Enrico Shippole[10], Kurt Bollacker[11], Tongshuang Wu[12], Luis Villa[13], Sandy Pentland[1] & Sara Hooker[14]

The race to train language models on vast, diverse and inconsistently documented datasets raises pressing legal and ethical concerns. To improve data transparency and understanding, we convene a multi-disciplinary effort between legal and machine learning experts to systematically audit and trace more than 1,800 text datasets. We develop tools and standards to trace the lineage of these datasets, including their source, creators, licences and subsequent use. Our landscape analysis highlights sharp divides in the composition and focus of data licenced for commercial use. Important categories including low-resource languages, creative tasks and new synthetic data all tend to be restrictively licenced. We observe frequent miscategorization of licences on popular dataset hosting sites, with licence omission rates of more than 70% and error rates of more than 50%. This highlights a crisis in misattribution and informed use of popular datasets driving many recent breakthroughs. Our analysis of data sources also explains the application of copyright law and fair use to finetuning data. As a contribution to continuing improvements in dataset transparency and responsible use, we release our audit, with an interactive user interface, the Data Provenance Explorer, to enable practitioners to trace and filter on data provenance for the most popular finetuning data collections: www.dataprovenance.org.

The latest wave of language models, both public[1–5] and proprietary[6–9] attribute their powerful abilities in large part to the diversity and richness of ever larger training datasets, including pretraining corpora, and finetuning datasets compiled by academics[10–12], synthetically generated by models[2,5] or aggregated by platforms such as Hugging Face[13]. Recent trends see practitioners combining and repackaging thousands of datasets and web sources[14–17], but despite some notable documentation efforts[18,19], there are diminishing efforts to attribute, document or understand the raw ingredients into new models[20–22].

A crisis in data transparency and its consequences

Increasingly, widely used dataset collections are being treated as monoliths, rather than a lineage of data sources, crawled (or model generated), curated and annotated, often with multiple rounds of repackaging (and relicensing) by successive practitioners. The disincentives to acknowledge this lineage stem both from the scale of modern data collection (the effort to properly attribute it), and increased copyright scrutiny[23]. Together, these factors have resulted in fewer datasheets[24], non-disclosure of training sources[6,7,25] and ultimately a decline in understanding training data[26,27].

A full list of affiliations appears at the end of the paper. ✉e-mail: rmahari@mit.edu

This lack of understanding can lead to data leakages between training and test data[28,29], expose personally identifiable information (PII)[30], present unintended biases or behaviours[31–33] and generally result in lower quality models than anticipated. Beyond these practical challenges, information gaps and documentation debt incur substantial ethical and legal risks. For instance, model releases appear to contradict data terms of use (for example, WizardCoder[34] licenced for commercial use, while training on commercially-prohibited OpenAI data), licence revisions postpublic release (with MPT-StoryTeller[35]) and even copyright lawsuits (for example, Andersen v. Stability AI[36] and Tremblay v. OpenAI[23]). As training models on data is both expensive and largely irreversible, these risks and challenges are not easily remedied. In this work, we term the combination of these indicators, including a dataset's sourcing, creation and licensing heritage, as well as its characteristics, the 'data provenance'.

Unreliable data provenance and licensing

Our work motivates the urgency of tooling that facilitates informed and responsible use of data in both pretraining and finetuning. To empower practitioners to attribute data provenance, we develop a set of tools and standards to trace the data lineage of 1,858 finetuning datasets from 44 of the most widely used and adopted text data collections. We compile and expand relevant metadata with a much richer taxonomy than Hugging Face, Papers with Code or other aggregators (see the 'DPExplorer' section). With legal experts, we design a pipeline for tracing dataset provenance, including the original source of the dataset, the associated licences, creators and subsequent use.

As a byproduct of our work establishing the data provenance of widely used datasets, we characterize the artificial intelligence (AI) data ecosystem and/or supply chain[37,38], and state of the field for policymakers, researchers and legal experts. Our work highlights a crisis in licence laundering and informed usage of popular datasets, with systemic problems in sparse, ambiguous or incorrect licence documentation. Notably, we find that more than 70% of licences for popular datasets on GitHub and Hugging Face are 'unspecified', leaving a substantial information gap that is difficult to navigate in terms of legal responsibility. The licences that are attached to datasets uploaded to dataset sharing platforms are often inconsistent with the licence ascribed by the original author of the dataset: our rigorous re-annotation of licences finds that 66% of analysed Hugging Face licences were in a different use category, often labelled as more permissive than the author's original licence. As a result, much of these data are risky to use (or harmfully misleading) for practitioners who want to respect author's intentions. Our initiative reduces unspecified licences from more than 72 to 30% and attaches licence URLs, allowing model developers to more confidently select appropriate data for their needs. To this end, the data provenance initiative supports attribution and responsible AI with the following contributions:

(1) The most extensive known public audit of AI data provenance, tracing the lineage of more than 1,800 text datasets (the 'DPCollection'), their licences, conditions and sources. We document changes in the dataset licensing landscape and synthesize observations into legal guidance for developers (see the 'Legal discussion' section).
(2) The Data Provenance Explorer (DPExplorer) (www.dataprovenance.org), an open-source repository for downloading, filtering and exploring data provenance and characteristics. Our tools auto-generate data provenance cards for scalable symbolic attribution and future documentation best practices.
(3) We find a sharp and widening divide between commercially open and closed data, with the latter monopolizing more diverse and creative sources. We suggest a data collection focus to narrow this gap.

The initiative to audit data provenance

The data provenance initiative's goal is to audit popular and widely used datasets with large-scale legal and AI expert-guided annotation. We propose a base set of indicators necessary for tracing dataset lineage and understanding dataset risks (described in the 'DPExplorer' section). As a first contribution of the initiative, we audit 44 instruction or 'alignment' finetuning data collections composed of 1,858 individual datasets, selected by experts for their widespread adoption and use in the community. The selected collections and their variants see hundreds to more than 10 million monthly downloads on Hugging Face, with the datasets within these collections tallying to many more (Table 1). While these metrics have limitations, especially for application-specific use cases, we hope that our reproducible pipeline will be extended to other datasets.

Our initiative's initial focus on alignment finetuning datasets was decided based on their growing emphasis in the community for improving helpfulness, reducing harmfulness and orienting models to human values[39]. Some collections have overlapping datasets and examples, but we choose not to deduplicate to preserve the original design choices, that may include different templates, formatting and filtering.

DPExplorer

Our information audit spans (1) identifier information, bridging metadata from several aggregators, including Hugging Face, GitHub, Papers with Code, Semantic Scholar and ArXiv, (2) detailed dataset characteristics for a richer understanding of training set composition and (3) dataset provenance for licensing and attribution. We expand our provenance metadata beyond just licences, because conversations with practitioners revealed they rely not only on data licences, but on a specific legal and ethical risk tolerance, parameterized by (i) the lineage of licences, (ii) the data source, (iii) the creator's identity and (iv) the precedence of adoption by other developers.

We release our extensive audit as two tools: (1) a data explorer interface, the DPExplorer for widespread use and (2) an accompanying repository for practitioners to download the data filtered for licence conditions. Practitioners are also able to generate a human-readable, markdown summary or data provenance card of the used datasets and compositional properties for languages, tasks and licences (see the 'Data provenance card as a data bibliography' section). Modern researchers training on hundreds of datasets often find it onerous to manually curate extensive data cards for these compilations[24,40]. We hope this tool will aid in writing the data attribution and composition sections of these documentation efforts, by providing auto-generated, copy-and-pastable dataframe summaries. Details on the collected data are provided in the 'Metadata details' section.

Licences in the wild

Based on our extensive study of empirical licence use for natural language processing (NLP) datasets, we identify a number of insights with relevance to practitioners and the wider community (see Extended Data Table 1 for a detailed breakdown). We note that this section treats datasets generated via OpenAI's services as subject to a 'non-commercial' use restriction, reflecting OpenAI's Terms of Use. However, these terms constitute a contractual agreement, not a copyright licence, potentially making them unenforceable against third parties who did not create the data using OpenAI (see the 'Legal discussion' section for a detailed discussion).

Frequency of licence types. Figure 1 shows the distribution of licences. The most common licences are CC-BY-SA 4.0 (15.7%), the OpenAI Terms of Use (12.3%) and CC-BY 4.0 (11.6%). We identify a long tail of licence variants with unique terms, and a large set of custom licences accounting for 9.6% of all recorded licences on their own. This wide licence diversity illustrates the challenge to startups and less

https://doi.org/10.1038/s42256-024-00878-8

Table 1 | Alignment tuning collections and their characteristics

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

Properties of the collections include the numbers of datasets, dialogues, unique tasks, languages, topics, text domains, Hugging Face monthly downloads ('Downs') and the average length of input and target text, by characters. The Source column indicates whether a collection includes human web text (W) or model-generated text (G). The dialogue formats of each collection can be: zero-shot (Z), few-shot (F), chain-of-thought (C), response ranking (R) and multi-turn dialogue (M). The Use column indicates whether a collection includes data licenced for commercial use (hatched circle▨), data with no licence (unspecified, grey circle●) and data only licenced for non-commercial or academic use (cross-hatched circle◈). Note that these licences are self-reported and their applicability is complicated, requiring legal consultation. The 'O' column indicates whether the collection includes OpenAI model generations, which may or may not affect commercial viability (see the 'Legal discussion' section)

resourced organizations attempting to navigate responsible training data collection, its legality and ethics.

Distribution of restrictive licences. In total, 85% of dataset licences request attribution, and 30% include a share-alike clause ('share alike' is a copyright term meaning adaptations or copies of a work must be released under the same licence as the original). Datasets that request attribution pose challenges for practitioners who commonly train on hundreds of datasets and either do not cite them at all[6,7,25] or simply cite an aggregation of data, which often falls short of the licence's attribution requirements. Furthermore, share-alike clauses pose challenges for practitioners repackaging data collections, usually when multiple conflicting share-alike licences are involved as there is no clear way to resolve them (such as Longpre et al.[17], Wang et al.[41] and others in the DPCollection). Frequently, practitioners will over-write share-alike licences with more restrictive or even less restrictive conditions.

Missing or unspecified licences. Investigating these involves comparing our manually reviewed licensing terms to the licences for the same datasets, as documented in the aggregators GitHub, Hugging Face and Papers with Code. Table 2 shows that these crowdsourced aggregators have an extremely high proportion of missing (unspecified) licences, ranging from 69 to 72%, compared to our protocol that yields only 30% unspecified. An unspecified licence leaves it unclear whether the

https://doi.org/10.1038/s42256-024-00878-8

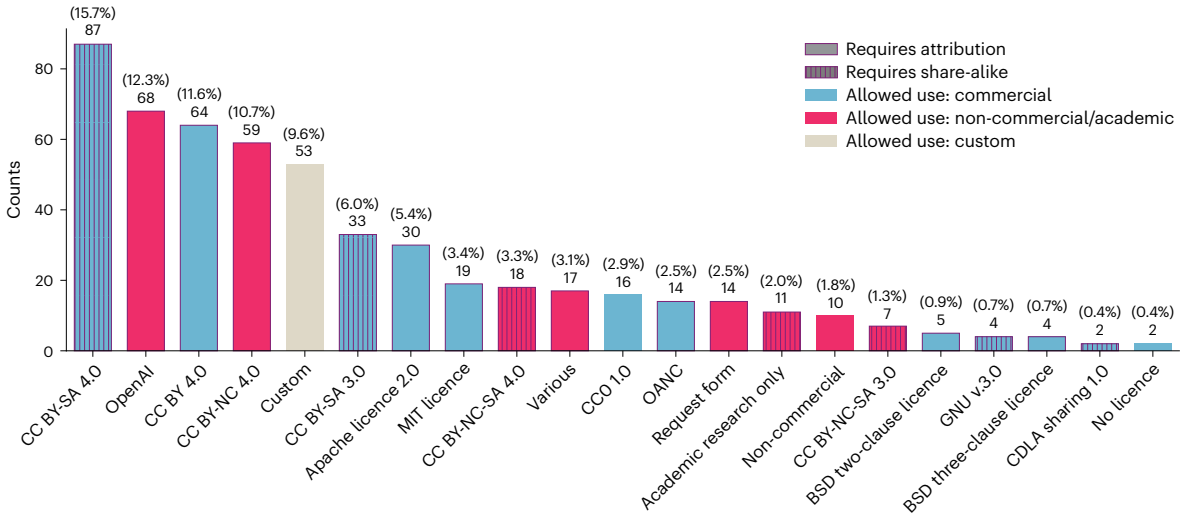


Fig. 1 | The distributions of licences used in the DPCollection, a popular sample of the major supervised NLP datasets. We find a long tail of custom licences, adopted from software for data: 73% of all licences require attribution and 33% share-alike, but the most popular are usually commercially permissive.

aggregator made a mistake or creators intentionally released data to the public domain. Consequently, risk-averse developers are forced to avoid many valuable datasets, which they would use if they were certain that there was no licence. As part of DPCollection, we manually reassign 46–65% of dataset licences (depending on the platform), resulting in much higher coverage, thus giving risk-averse developers more confidence and breadth in their dataset use.

Incorrectly specified licences. Table 2 shows that correct licences are frequently more restrictive than the ones by assigned by aggregators. GitHub, Hugging Face and Papers with Code each label licence use cases too permissively in 29%, 27% and 16% of cases, respectively. Our inspection suggests this is due to contributors on these platforms often mistaking licences attached to code in GitHub repositories for licences attached to data.

How does data availability differ by licence use category?

While non-commercial and academic-only licences play important roles in protecting data use, their presence can also exclude communities from participating (or competing) in the development of these technologies. In this section, we break down datasets according to their licence restrictions and see how they differ. Specifically, we ask: does complying with licences dictate systematic differences in resources for commercially permissive ('open') and non-commercial ('closed') development? And what particular features of data are particularly constrained by non-commercial prohibitions?

We compare datasets by categories of permitted use, according to their licences: (1) commercially viable, (2) non-commercial/academic-only (NC/A-O) or (3) unspecified licence. We group together non-commercial and academic-only conditions as the distinction plays a minor role in practice. We argue in the 'Legal discussion' section that datasets without any licence (unspecified) do not impose any conditions, may be treated as commercially viable, but this assessment depends on a developer's risk tolerance and jurisdiction.

Non-commercial and academic-only licensed datasets have greater diversity in tasks, topics, sources and target text lengths. For each of these features, Table 3 illustrates the mean number per dataset, broken down by licence category and entropy to measure the randomness, and thus diversity, of each feature. NC/A-O datasets see greater diversity of tasks, topics and sources represented in the text than commercial datasets. Extended Data Fig. 2 shows where this diversity comes from. The most NC/A-O task categories include brainstorming, explanation, logic and maths, as well as creativity and creative writing. In comparison, the most commercially viable task categories are short text generation, translation and classification. Similarly, among source domains, governments and search queries are largely viable for commercial (and unspecified) purposes, whereas general web, exams and model-generated sources are among the most restrictive.

Target text lengths are notably longer for NC/A-O datasets. Not only do NC/A-O datasets appear more textually and functionally diverse, their length characteristics differ substantially. While Table 3 shows the input text lengths across licence categories are similar on average, the target text lengths are higher for NC/A-O datasets (103 versus 677). This breakdown is further illustrated in Fig. 2, where we see greater representation of both NC/A-O and synthetic datasets above the 100 target token threshold (*y* axis).

The rise of synthetic datasets generated using APIs with non-commercial terms of use may explain the differences in text diversity and length. Table 3 also shows a full 45% of NC/A-O datasets are synthetic, compared to <14% in more permissive licence categories. Taori et al.[2], Wang et al.[5], Touvron et al.[4], Xu et al.[42] and their variants, all generated in part using commercial APIs, exhibit stronger task and topic diversity than traditional academic datasets, as they cater to longer form generations, by design. This is evident from the concentration of creative, brainstorming and reasoning tasks baked into them, compared to the focus of more topic-focused question answering, classification and short text generation in non-synthetic datasets. These datasets are usually created using larger proprietary models, mostly from OpenAI APIs (see the 'Legal discussion' section).

In 2023 there was a spike in NC/A-O dataset licences. Among the large collection of datasets we trace, we record the date at which they are released, by cross-referencing their associated GitHub, ArXiv and Hugging Face dates. We find a striking change in the pattern of licensing restrictions. As shown in Extended Data Fig. 1, before 2023, no year saw more than one-third of the datasets released as NC/A-O. However, in 2023, when many of the most popular and diverse datasets were published, the NC/A-O rate is 61%. Furthermore, most datasets were unaccompanied by a licence before 2022 (~50–80%), compared to only 12% in 2023. The shift to more licence use, and to more restrictive licences, may foreshadow future challenges to open data.

Table 2 | The distribution of licence use categories shows our licences have far fewer unspecified omissions than GitHub (GH, 72%), Hugging Face (HF, 69%) and Papers with Code (PWC, 70%), categorizing licences more confidently into commercial or non-commercial categories

Correct licence		Licence according to aggregators				
Licence	Count	Aggregators	Commercial	Unspecified	Non-commercial	Academic only
Commercial	856 (46.1%)	GH	349	507	0	0
		HF	176	677	1	2
		PWC	313	520	1	22
Unspecified	570 (30.7%)	GH	112	458	0	0
		HF	164	395	6	5
		PWC	31	523	1	15
Non-commercial	352 (19.0%)	GH	49	303	0	0
		HF	113	152	80	7
		PWC	2	191	157	2
Academic-only	80 (4.3%)	GH	9	71	0	0
		HF	9	65	2	4
		PWC	5	65	2	8
Total	1,858 (100%)	GH	519 (28%)	1,339 (72%)	0 (0%)	0 (0%)
		HF	462 (25%)	1,289 (69%)	89 (5%)	18 (1%)
		PWC	351 (19%)	1,299 (70%)	161 (9%)	47 (3%)

GitHub, Hugging Face and Papers with Code match our licences (grey regions) 43, 35 and 54% of the time, respectively, and suggest incorrect licences that are too permissive 29, 27 and 16% of the time.

Table 3 | The mean number of features (for example, tasks or languages) per dataset, and the mean entropy of the distribution, representing the diversity of categories

Metrics	Commerical		Unspecified		NC/A-O	
	Mean	Entropy	Mean	Entropy	Mean	Entropy
Tasks	1.7±0.1	0.61	1.6±0.1	0.53	3.4±0.2	0.69
Languages	1.3±0	0.52	1.2±0	0.16	1.1±0	0.45
Topics	8.2±0.2	0.70	9.2±0.1	0.75	9.1±0.2	0.77
Sources	1.6±0.1	0.67	1.8±0.1	0.72	4.2±1.3	0.78
Input target lengths	1,043.4±151.9	6.37	860.2±67.7	6.66	950.3±112.9	6.46
Target text lengths	102.7±14.6	4.39	90.5±14.3	4.09	1,580.7±965.6	5.37
Synthetic	12.8±2.1	-	13.6±1.7	-	45.5%±3.4	-

Non-commercial and academic-only datasets have consistently and statistically higher task, topic and source variety than commercial datasets. We use normalized Shannon entropy for discrete features and differential entropy for continuous features, which are both measures of randomness.

Commercial datasets have greater language variety, but low-resource language datasets see the least commercial coverage. Table 3 shows that commercial datasets have greater diversity of languages than NC/A-O. However, when broken down by language family, as in Extended Data Fig. 1, we see stark differences in permitted use by group. Code language datasets are nearly all commercially viable (78%), because dataset creators can easily filter GitHub for permissively licenced repositories. English, Atlantic-Congo and Afroasiatic languages also see large permissive representation. However, Turkic, Sino-Tibetan, Japonic and Indo-European languages see in excess of 35% as non-commercial. Note that while the Indo-European language family contains many high-resource European language families, there is a long tail of lower-resource ones. These NC/A-O language families provide directions for open data practitioners to focus their future efforts.

Broader characteristics of the data

In addition to understanding systematic differences in the data by licence, there are research questions regarding the overall composition and characteristics of these widely used and adopted datasets. Our compilation of metadata through the DPCollection allows us to map the landscape of data characteristics and inspect particular features. Note that all these details are also available with interactive visualizations at www.dataprovenance.org, for further research and examination.

Language representation is heavily skewed to English and western European languages. Following Talat et al.'s[43] recommendations in data transparency and documentation in demographic analysis, and corroborating Kreutzer et al.'s[44] similar analysis for pretraining corpora, we find a stark Western-centric skew in representation. Figure 3 illustrates the coverage per country according to the spoken languages and their representation in DPCollection (see Methods for details). Figure 3 shows that Asian, African and South American nations are sparsely covered if at all. Even when nations from the Global South appear to have linguistic representation, the text source and dialect of the language contained in these datasets almost always originates from North American or European creators and web sources (although this is difficult to measure precisely). These observations corroborate similar findings in

https://doi.org/10.1038/s42256-024-00878-8

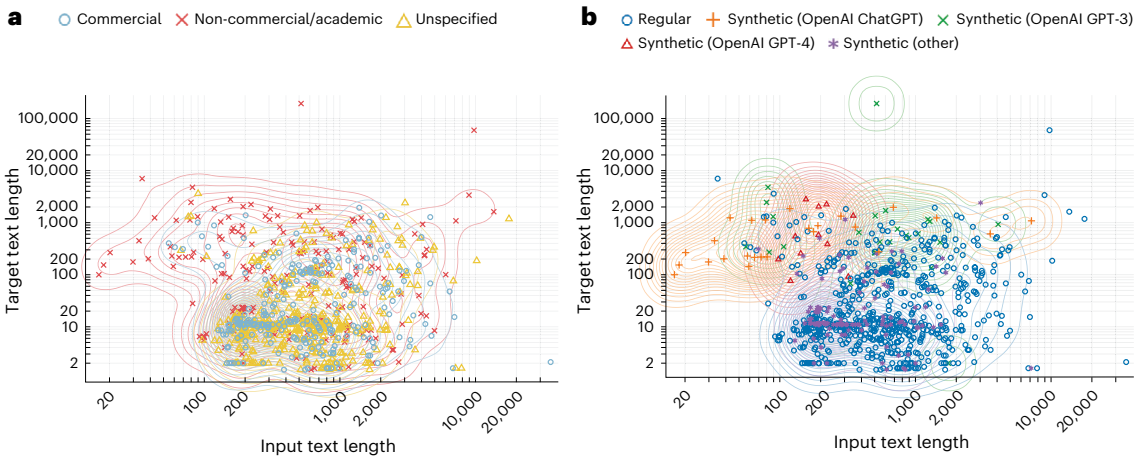


Fig. 2 | Across finetuning datasets, we visualize their mean input and target text lengths, measured in log-scaled number of characters. The colours indicate either their licence use category (left) or whether they were machine generated or human collected (right). Long target texts are represented in large part by non-commercial and synthetic datasets that are often generated by commercial APIs. **a**, Licence use categories versus text lengths (log-scaled character length). **b**, Synthetic and/or regular datasets versus text lengths (log-scaled character length).

the geo-diversity of image data in the vision domain[45–47]. Models trained on these datasets are likely to have inherent bias, underperforming in critical ways for users of models outside the west[48].

The primary drivers of dataset curation are academic organizations, industry labs, and research institutions. These metrics describe the scale of dataset curation contributions, but not the influence each dataset has had on the community. Extended Data Table 1a demonstrates the single largest dataset contributors are AI2 (12.3%), University of Washington (8.9%) and Facebook AI Research (8.4%). It is important to note that these contributors often only download and compile text from the Internet that was originally written by other people. Most dataset creators are located in the United States and China, raising additional concerns about potential biases contained in lower-resource language datasets.

Text datasets focus on language topics, general knowledge, logic and lifestyle. Previous data collection work focuses predominantly on describing datasets by their task compositions[5,11,17], but rarely by their actual topics (except ref. 14 in their appendix). Extended Data Table 1b shows the most popular topics, clustered by category, with their representation across datasets. Like most NLP tasks, much of these text data focus on communication and language understanding topics, followed closely by general knowledge, routine, sports and education.

Text datasets are sourced primarily from online encyclopaedias, social media, and the web. While practitioners document their individual dataset sources in their published papers, this information is unstructured and can be hard to find. Collection of widely used datasets commonly just cite data papers rather than their sources, and data sources are often lost during data compilation and repackaging. By manually scanning approximately 500 academic papers, we annotate the original text sources and compile them into domain clusters to permit attribution and analysis, as summarized in Extended Data Table 1c. Among the most widely used sources are wikipedia.org (14.9%), undisclosed webpage crawls (7.0%), Reddit (6.2%) and Twitter (4.0%). The least represented domains include commerce, reviews, legal, academic papers and search queries.

Legal discussion

Our empirical analysis highlights that we are in the midst of a crisis in dataset provenance and practitioners are forced to make decisions based on limited information and opaque legal frameworks. While we believe our tooling will enable better transparency about where licences are in tension, major legal ambiguities remain in data licensing.

Open legal question regarding copyright and model training

Apart from the jurisdictional and interpretive ambiguities discussed in the Supplementary Information Legal Discussion, the process of training a model raises specific copyright questions[49]. Training a model poses several interesting legal questions with respect to copyright and infringement may occur in several ways even before any outputs are generated. First, the act of creating a training dataset by crawling existing works involves making a digital copy of the underlying data. As the name implies, copyright gives the author of a protected work the exclusive right to make copies of that work (17 US Code § 106). If the crawled data is protected by copyright, then creating training data corpora may raise copyright issues[50]. Second, copyright holders generally have an exclusive right to create derivative works (for example, translations of a work). Should a trained machine learning model be considered a derivative of the training data[51]? If so, then training a model would be more likely to violate the rights of the training data's copyright holders[52].

In the United States, the fair use exception may allow models to be trained on protected works (17 US Code § 107)[53–56]. As explained by previous work, the training of machine learning models on copyrighted content may be permissible if the underlying works are sufficiently 'transformed' into model weights, only a small amount of each work in the training data is included in the trained model, model training is designed to only glean generalizable insights from the training data, and the trained model does not have a strong effect on the economic success of the works in the training data. It is important to underscore that, while training a machine learning model itself may be protected by fair use this does not mean that model outputs will not infringe on the copyright of previous works. As the authors above highlight, the application of fair use in this context is still evolving and several of these issues are currently being litigated (for example, Andersen v. Stability[36], Doe v. GitHub[57] and Tremblay v. OpenAI[23]).

Fair use for data created for machine learning

Fair use is less likely to apply when works are created for the sole purpose of training machine learning models as in the case of supervised datasets with copyrightable compositions or annotations. Most literature on fair use and machine learning focuses on copyrighted art or text

https://doi.org/10.1038/s42256-024-00878-8

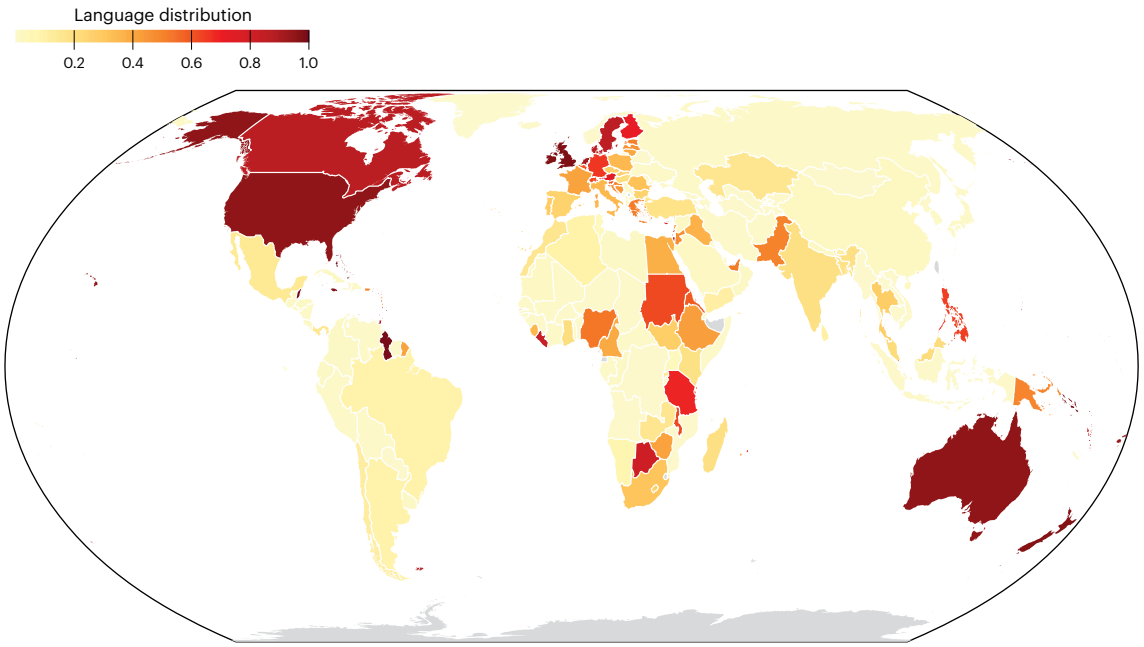


Fig. 3 | A global heatmap of language representation scores measuring how well each country's spoken languages are represented by the composition of natural language datasets in DPCollection, as calculated in the 'Computing language representation' section. English-speaking and western European nations are best represented, while the Global South sees limited coverage.

that was crawled to train a model. These crawled works were not created for the purpose of training machine learning models. By contrast, in this paper, we focus on supervised datasets that were created for the sole purpose of training machine learning models. As underscored by refs. 53 and 55, the fair use analysis depends in part on whether a trained model copies the 'expressive purpose' of the original work (Bill Graham Archives v. Dorling Kindersley[58]). While the expressive purpose of a piece of text or art is not to train machine learning models, the purpose of a training dataset is to do just that. As a result, we expect that it is less likely that fair use would apply to the use of curated data. Instead, the creators of these datasets hold a copyright in the dataset and the terms of the dataset licence agreement govern the subsequent use of these data. However, it is rare in practice for a large language model (LLM) to use a single supervised dataset and often multiple datasets are compiled into collections. This further complicates the legal analysis because we find that the licence terms of many popular dataset collections are conflicting.

Legal implications of LLM-generated annotations

We find that approximately 12% of the datasets we audit were annotated using OpenAI. The OpenAI Terms of Use state that outputs from the OpenAI service may not be used to 'to develop models that compete with OpenAI' (https://openai.com/policies/terms-of-use). These terms seem to preclude a developer from using OpenAI to generate training data to train a competing LLM. However, it is not clear whether they would also limit the ability of a developer to use OpenAI to create and publish an annotated dataset. While publishing such a dataset does not directly compete with OpenAI, it seems foreseeable that such a dataset could enable third parties (who did not themselves use OpenAI) to create competing LLMs. In the United States, there are several doctrines of secondary or indirect copyright liability aimed to enforce copyright in cases where there is no direct infringement[51,59]. The application of these doctrines depends on many factors, most importantly on whether OpenAI has a copyright interest in its outputs. If these copyright doctrines do not apply, then it is still possible that publishing the dataset constitutes a breach of contract by the dataset developers. While it would be more challenging for OpenAI to pursue a case against third parties, there are myriad other business torts, from unfair competition to misappropriation, that may be relevant to this situation and which go beyond the scope of this paper[60]. Time will tell whether OpenAI and other LLM providers can enforce their terms against third parties. However, a prominent researcher at Google has already resigned citing concerns that OpenAI outputs were used to train BARD[61]. In light of these ambiguities, our tool gives developers the ability to exclude OpenAI-generated datasets.

Data provenance enables informed decision-making

Despite these pervasive legal uncertainties, practitioners can still make some informed decisions to minimize risk if they have reliable data provenance information. With access to this information, practitioners can decide to err on the side of caution and to use only data licenced for commercial use, contact dataset creators of restrictively licenced data to negotiate a usage agreement or decide that their specific context and risk tolerance allows them to use datasets licenced for non-commercial use. Through our audit and tooling, we seek to provide the information needed to make informed decisions in an otherwise ambiguous landscape. Model providers may also consider strategies for partially mitigating uncertainties for downstream users, for example, by indemnifying users, as done by Google Cloud[62]. Of course, this does not solve the issues faced by model developers or dataset curators. We urge practitioners to take dataset licences seriously, as they may have real impacts on how their models may be used in practice.

In creating a repository of data licensing information, we hope to also encourage dataset creators to be more thoughtful about the licences that they select. Dataset creators are well-positioned to understand the appropriate uses of the datasets they publish and licences can be a tool to communicate these restrictions and to encourage responsible AI development.

Finally, this discussion highlights an important opportunity for regulators to reduce legal ambiguity by clarifying the enforceability of dataset licences both to help catalyse innovation and as a way to promote more responsible, inclusive and transparent machine learning practices[63,64].

Methods

Details on collecting data provenance

These data were collected with a mix of manual and automated techniques, leveraging dataset aggregators such as GitHub, Hugging Face and Semantic Scholar (Extended Data Fig. 3). Annotating and verifying licence information, in particular, required a carefully guided manual workflow, designed with legal practitioners ('License annotation process' section). Once these information aggregators were connected, it was possible to synthesize or crawl additional metadata, such as dataset languages, task categories and time of collection. And for richer details on each dataset, such as text topics and source, we used carefully tuned prompts on language models inspecting each dataset.

Automated annotation methods. Based on the manually retrieved pages, we automatically extract licences from Hugging Face configurations and GitHub pages. We leverage the Semantic Scholar public API[65] to retrieve the released date and current citation counts associated with academic publications. Additionally, we compute a series of other helpful, but often overlooked data properties such as text metrics (the minimum, mean and maximum for input and target lengths) and dialogue turns. We elected to measure sequence length in characters rather than word tokens, for fairer treatment across language and script given well-known differences in tokenizer performance across different languages[66].

API annotation methods. While task categories have become the established measurement of data diversity in recent instruction tuning work[5,11], there are so many other rich features describing data diversity and representation. To augment this, we use OpenAI's GPT-4 API to help annotate for text topics. We randomly sampled 100 examples per dataset and carefully prompt GPT-4 to suggest up to ten topics discussed in the text.

To annotate for the original data sources, AI experts (PhD students and postdocs) reviewed the papers and filled out the original text sources, whether machines or template-generation were used for synthetic generation, and whether human annotators were used. GPT-4 was used as an in-context retriever on the dataset's ArXiv paper to extract snippets that the experts may have missed. We split the ArXiv paper into 4,000-character chunks and prompt the API to return a json list of any mentions of the dataset source, for example from crawling, synthetic or manual generation.

Licence annotation process

One of our central contributions is to validate the licences associated with widely used and adopted datasets. This process provides a current snapshot of the data provenance landscape for finetuning data, but the methods and code we develop and share here are aimed to facilitate future audits, including those that extend beyond finetuning and text data. This followed a time-intensive human annotation protocol to collect dataset authors' self-reported licences and categorize them according to stated conditions. Note that this protocol reflects best efforts to verify self-reported licences and does not constitute legal advice. Additionally, it is important to note that the enforceability of these licences depends on several factors discussed in the 'Legal discussion' section. One especially important assumption in cases where datasets are based on data obtained from other sources is that dataset creators actually have a copyright interest in their dataset. This depends on the data source and how creators modify or augment these data, and requires a case-by-case analysis. However, it appears that most developers operate under the general assumption that they alone own their datasets. Our licence annotation workflow follows these steps:

(1) Compile all self-reported licence information. We aggregate all licensing information reported on GitHub, ArXiv, Hugging Face, Papers with Code and the collection itself (for example, Super-Natural Instructions)[41].
(2) Search for explicit data licences. The annotator searches for a licence specifically given to the dataset (not the accompanying code) by the authors. A licence is found if (i) the GitHub repository mentions or links a licence in reference to the data, (ii) the Hugging Face licence label was uploaded by the dataset creator themselves or (iii) the paper, Hugging Face or Papers with Code provide a dataset-specific licence link, attributable to the data authors.
(3) Identify a licence type. A licence may fall into a set of common types (for example, MIT, Apache 2, CC BY SA and so on), be a 'Custom' licence, a permission request form or, if none was found for the data, unspecified. If a dataset has multiple licences, the annotator will list each of them according to their types.
(4) Categorize licences. From the perspective of a machine learning practitioner, licensing typically is viewed through the lens of how it affects the model lifecycle—does it impede or allow for training on the data, downstream use conditions, attributing, modifying or re-distributing it? On the basis of discussions with industry experts, we categorize licences based on three important questions that affect the model lifecycle: is data usage limited to academic or non-commercial purposes (permitted use), does the data source need to be attributed (attribution) and do derivatives of the data need to be licenced under the same terms as the original (share-alike)? If there are multiple licences for a dataset, its categorization for each feature is chosen as the strictest across licences.
(5) Sources. For each dataset, we review the documentation available in the academic paper, GitHub, website or Hugging Face to determine the original sources of the text as precisely as possible. The original sources are where the text was taken from before it was used in datasets. Sometimes, a dataset (introduced in a specific paper) might be based on another dataset. For example, the dataset might be an extension of another dataset, or it could be taking one dataset and formatting and/or modifying it to be usable for another learning task. In these cases, we find the 'root' dataset (that is, the original one that is extended or modified) and determine what the source is for that particular dataset. We also include new text sources that have been leveraged at each stage of dataset derivation and development. We provide a list of sources, grouped by domain, at https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection/blob/main/constants/domain_groups.json.
(6) Additional provenance. In practice, legal teams may wish to balance their risk tolerance with more nuanced criteria. For instance, they may be satisfied with using (more permissive) GitHub licences, even when it is ambiguous whether these apply to the code or the data. They may also wish to include or exclude datasets on the basis of whether these are already widely used in practice, where the original data were sourced from and if the creator is a competitor. To supplement the above licence categories, we also collect all this metadata for fine-grained selection and filtering.

Data provenance card as a data bibliography

Previous work has stressed the importance of data documentation and attribution[22,67]. In particular, Gebru et al.'s[24] datasheets break down documentation into motivation, composition, collection process, processing, uses, maintenance and distribution. Similarly, Bender and Friedman[67] ask for curation rationale, language variety, speaker demographic, annotator demographic, speech situation and text characteristics, among others. However, when models train on many sources of data, even if they are each rigorously documented for each

of these fields (rarely the case), it is challenging to cleanly synthesize comprehensive and navigable documentation for the resulting bundle.

To make this process tractable with scale, we propose leveraging symbolic attribution, where our tools auto-generate a structured store of the provenance and attribution metadata, similar to a bibliography for data (these are auto-generated at https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection). Our collected schema allows this store to succinctly capture the attribution (links to repositories, aggregator copies, papers, creators), provenance (text/machine sources, licences) and compositional properties of the data (languages, tasks, text metrics, format and time). This file of references and metadata, known as a data provenance card, enables comprehensive documentation proposed by previous work while providing some advantages from its structure. First, the data provenance card can be easily searched, sorted, filtered and analysed, whereas datasheets or statements, designed for individual datasets, are meant to be manually read. Second, developers can efficiently assemble relevant information without losing any detail by symbolically linking to the original datasets and their documentation. Third, as datasets are continually repackaged and absorbed into newer and bigger collections, data provenance cards are easily adaptable by simply appending or concatenating them. Altogether, we hope this tooling enables and promotes the thorough documentation proposed in previous work[24,40,67,68]

Metadata details

Collecting comprehensive metadata for each dataset required leveraging several sources including collection by linking to resources already on the web (W), human annotation by legal experts (E) or using GPT-4 to assist in human annotation (G). The collected metadata cover many aspects of these datasets, spanning identifiers, dataset characteristics and provenance information. These features were selected on the basis of our input from machine learning experts who contributed to this paper and who identified the information that would be most useful to practitioners.

Identifier information. Identifier information discloses links and connects aggregator identifiers.

(1) Dataset identifiers (E): the dataset's name, associated paper title and description of the dataset.
(2) Dataset aggregator links (E): a link to each major aggregator, including GitHub, Hugging Face, Papers with Code, Semantic Scholar and ArXiv, allows us to incorporate and compare their crowdsourced metadata.
(3) Collection (E): the name and URL to the data collection of which this dataset is a part.

Dataset characteristics. Dataset characteristics are detailed information relevant to understanding data representation and/or composition, and curating a training set.

(1) Languages (E): each of the languages represented in the dataset, so developers can easily follow the 'bender rule'[69].
(2) Task categories (E, G): the 20+ task categories represented in the instructions, such as question answering, translation, programme synthesis, toxicity identification, creative writing and roleplaying.
(3) Text topics (G): an automated annotation of the topics discussed in the datasets, with GPT-4 labelling a sample of 100 examples for up to ten covered topics.
(4) Text length metrics: the minimum, maximum and mean number of dialogue turns per conversation of characters (agnostic to tokenization/non-whitespace languages, as this introduces biases[66]) per user instruction and assistant responses.
(5) Format (E): the format and intended use of the data. The options are zero-shot prompts, few-shot prompts, chain-of-thought prompts, multi-turn dialogue and response ranking.
(6) Time of collection (W): the time when the work was published, which acts as an upper bound estimate of the age of the text.

Dataset provenance.

(1) Licences (W, E): the licence name and URLs associated with the data, using the process described in the 'Licence annotation process'. We also enable filtering by licence use classes categorized by legal professionals.
(2) Text source (E, G): the original sources of the text, often Wikipedia, Reddit or other crawled online or offline sources.
(3) Creators (E): the institutions of the dataset authors, including universities, corporations and other organizations.
(4) Attribution (W): the attribution information for the authors of the paper associated with the dataset.
(5) Citation and download counts (W): the citation and Hugging Face download count for the paper and dataset, dated September 2023. This acts as an estimate of community use, and is commonly used as precedence to decide on the risk level for using these datasets.

Developing the DPExporer

The DPExplorer displays the collected data in a format accessible to developers by applying different aggregation, specialized filtering and tallying steps to obtain data summary statistics and overviews. All plots are built in JavaScript using the observablehq, P5 and D3 libraries that support dynamic, interactive visualizations. Many of our plots visualize languages and creators across geographies. To situate these, we use lookup tables, such as the language ISO 639 to group language families and we use the topojson to visualize the world map. We also map those to country codes and to language codes to interface with the map. As done in this paper, we map all tasks, topics and licences into clustered categories (Extended Data Table 2) to allow us to plot their distributions. We manually predefine clusters based on discussion among the authors, frequent taxonomies already used in the field, coupled with manual observation and iteration for what was tractable.

Computing language representation

We compute a language representation score S_k for each country k, parametrized by p_{kl}, the percentage of people in country k that speak language l, and w_{li} that is a binary indicator of 1 if dataset $i \in D$ contains language l and 0 otherwise.

$$S_k = \sum_{l \in L} \left(p_{kl} \times \sum_{i \in D} w_{li} \right)$$

Software

We use the following Python (v.3.8.9) packages: aiohttp (v.3.9.5), aiosignal (v.1.3.1), annotated-types (v.0.7.0), anyio (v.4.4.0), async-timeout (v.4.0.3), attrs (v.23.2.0), certifi (v.2023.7.22), chardet (v.5.2.0), charset-normalizer (v.3.3.2), ConfigArgParse (v.1.7), datasets (v.2.19.2), dill (v.0.3.8), distlib (v.0.3.6), distro (v.1.9.0), exceptiongroup (v.1.2.1), filelock (v.3.11.0), frozenlist (v.1.4.1), fsspec (v.2024.3.1), h11 (v.0.14.0), httpcore (v.1.0.5), httpx (v.0.27.0), huggingface-hub (v.0.23.3), idna (v.3.4), jsonlines (v.4.0.0), multidict (v.6.0.5), multiprocess (v.0.70.16), numpy (v.1.24.4), openai (v.1.33.0), packaging (v.24.1), pandas (v.2.0.3), platformdirs (v.3.2.0), pyarrow (v.16.1.0), pyarrow-hotfix (v.0.6), pydantic (v.2.7.3), pydantic_core (v.2.18.4), python-dateutil (v.2.9.0.post0), python-dotenv (v.1.0.1), pytz (v.2024.1), PyYAML (v.6.0.1), requests (v.2.32.3), semanticscholar (v.0.5.0), sniffio (v.1.3.1), tabulate (v.0.9.0), tenacity (v.8.2.3), tqdm (v.4.66.4), typing_extensions (v.4.12.2), tzdata (v.2024.1), urllib3 (v.2.1.0), virtualenv (v.20.21.0), xxhash (v.3.4.1) and yarl (v.1.9.4).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in our analysis, including the manually collected data, as well as a generalizable pipeline for future data collection, can be found in our public repository: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection. Extended Data Table 1 summarizes the data sources for our work and a full list of data sources may be found at: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection/tree/main/data_summaries. These repositories contain all the metadata we collected and build out downloaders that pull from Hugging Face or GitHub to standardize formats, wrap them in their metadata and then apply tools to filter, sort, select and visualize those datasets. From these collections, we identify text datasets for multi-task finetuning, preference and/or human feedback tuning and multi-turn dialogue. These are selected by compiling popular datasets on Hugging Face, for a diverse set of tasks, as well as other popular datasets we discovered in the process of investigating popular instruction tuned models on Hugging Face for general-purpose chatting, tool-use, multilingual questions and answers, and other common NLP tasks. Although this process is partly subjective, we devise an annotation pipeline (described in the 'Metadata details' section) to maximize reproducibility. The annotated data may be accessed, visualized and explored on https://dataprovenance.org/.

Code availability

All code used for our analysis and to produce figures may be found in our GitHub repository[70]. The code used to develop the DPExplorer is available at: https://github.com/shayne-longpre/opal-dl-streamlit. We provide this example notebook to show how we generate our visualizations: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection/blob/main/src/analysis/text_ft_plots.ipynb. Our data analysis and collection pipeline included both manual collection and automated data preparation and/or analysis using latest standard libraries at the time of submission.

References

1. Chung, H.W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).
2. Taori, R. et al. Stanford alpaca: an instruction-following Llama model. *GitHub* https://crfm.stanford.edu/2023/03/13/alpaca.html (2023).
3. Geng, X. et al. Koala: a dialogue model for academic research. *Berkeley Artificial Intelligence Research* https://bair.berkeley.edu/blog/2023/04/03/koala/ (2023).
4. Touvron, H. et al. Llama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
5. Wang, Y. et al. Self-instruct: aligning language model with self generated instructions. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Rogers, A. et al.) 13484–13508 (Association for Computational Linguistics, 2023).
6. Anil, R. et al. Palm 2 technical report. Preprint at https://arxiv.org/abs/2305.10403 (2023).
7. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
8. Model card and evaluations for Claude models. *Anthropic* https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf (Anthropic, 2023).
9. Yoo, J., Perlin, K., Kamalakara, S. R. & Araújo J. G. Scalable training of language models using JAX-pjit and TPUv4. Preprint at https://arxiv.org/abs/2204.06514 (2022).
10. Wei, J. et al. Finetuned language models are zero-shot learners. In *Proc. 2022 International Conference on Learning Representations* https://openreview.net/pdf?id=gEZrGCozdqR (ICLR, 2022).
11. Sanh, V. et al. Multitask prompted training enables zero-shot task generalization. In *Proc. 2022 International Conference on Learning Representations* https://openreview.net/pdf?id=9Vrb9D0WI4 (ICLR, 2022).
12. Muennighoff, N. et al. Crosslingual generalization through multitask finetuning. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Rogers, A. et al.) 15991–16111 (Association for Computational Linguistics, 2023).
13. Lhoest, Q. et al. Datasets: a community library for natural language processing. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Adel, H. & Shi, S.)175–184 (Association for Computational Linguistics, 2021).
14. Gao, L. et al. The pile: an 800 GB dataset of diverse text for language modeling. Preprint at https://arxiv.org/abs/2101.00027 (2020).
15. Penedo, G. et al. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. In *Proc. of the 37th International Conference on Neural Information Processing Systems* 79155–79172 (Curran, 2023)
16. Wang, Y. et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y. et al.) 5085–5109 (Association for Computational Linguistics, 2022).
17. Longpre, S. et al. The flan collection: designing data and methods for effective instruction tuning. In *Proc. of the 40th International Conference on Machine Learning* https://openreview.net/pdf?id=ZX4uS605XV (2023).
18. Gaia search tool https://huggingface.co/spaces/spacerini/gaia (Spacerini, 2021).
19. Biderman, S., Bicheno, K. & Gao, L. Datasheet for the pile. Preprint at https://arxiv.org/abs/2201.07311 (2022).
20. Dodge, J. et al. Documenting large webtext corpora: a case study on the colossal clean crawled corpus. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Adel, H. & Shi, S.) 1286–1305 (Association for Computational Linguistics, 2021).
21. Bandy, J. & Vincent, N. Addressing 'documentation debt' in machine learning research: a retrospective datasheet for bookcorpus. In *Proc. of the Neural Information Processing Systems Track on Datasets and Benchmarks* (eds Vanschoren, J. & Yeung. S.) https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/54229abfcfa5649e7003b83dd4755294-Paper-round1.pdf (2021).
22. Bommasani, R. et al. The foundation model transparency index. Preprint at https://arxiv.org/abs/2310.12941 (2023).
23. Tremblay v. OpenAI, Inc., 3:23-cv-03223-AMO (N.D. Cal. 2024).
24. Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
25. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
26. Sambasivan, N. et al. 'Everyone wants to do the model work, not the data work': data cascades in high-stakes AI. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (eds Kitamura, Y. & Quigley, A.) https://doi.org/10.1145/3411764.34455 (ACM, 2021).
27. Longpre, S. et al. A pretrainer's guide to training data: measuring the effects of data age, domain coverage, quality, & toxicity. In *Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (eds Duh, K. et al.) 3245–3276 (Association for Computational Linguistics, 2024).

28. Elangovan, A., He, J. & Verspoor, K. Memorization vs. generalization: quantifying data leakage in NLP performance evaluation. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (eds Merlo, P. et al.) 1325–1335 (ACM, 2021).
29. Carlini, N. et al. Quantifying memorization across neural language models. In *Proc. 2023 International Conference on Learning Representations* https://openreview.net/pdf?id=TatRHT_1cK (ICLR, 2023).
30. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with gpt-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
31. Welbl, J. et al. Challenges in detoxifying language models. In *Proc. Findings of the Association for Computational Linguistics: EMNLP 2021* (eds Moens, M.-F. et al.) 2447–2469 (ACM, 2021).
32. Xu, A. et al. Detoxifying language models risks marginalizing minority voices. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Toutanova, K. et al.) 2390–2397 (ACM, 2021).
33. Pozzobon, L., Ermis, B., Lewis, P. & Hooker, S. On the challenges of using black-box APIs for toxicity evaluation in research. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 7595–7609 (Association for Computational Linguistics, 2023).
34. Luo, Z. et al. Wizardcoder: empowering code large language models with evol-instruct. In *Proc. 12th International Conference on Learning Representations* https://openreview.net/pdf?id=UnUwSIgK5W (ICLR, 2024).
35. Frankle, J. Tweet by mosaic ML. *Twitter* https://twitter.com/jefrankle/status/1654848529834078208 (2023).
36. Andersen v. Stability AI Ltd., 23-cv-00201-WHO (N.D. Cal. 2023).
37. Cen, S. H. et al. AI supply chains (and why they matter). The second post in our series On AI Deployment. *Substack* https://aipolicy.substack.com/p/supply-chains-2 (2023).
38. Bommasani, R., Soylu, D., Liao, T. I., Creel, K. A. & Liang, P. Ecosystem graphs: the social footprint of foundation models. Preprint at https://arxiv.org/abs/2303.15772 (2023).
39. Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Proc. of the 36th International Conference on Neural Information Processing Systems* 27730–27744 (Curran, 2024).
40. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency* 220–229 (ACM, 2019).
41. Wang, Y. et al. Super-natural instructions: generalization via declarative instructions on 1600+ NLP tasks. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y. et a.) 5085–5109 (Association for Computational Linguistics 2022).
42. Xu, C. et al. WizardLM: empowering large language models to follow complex instructions. In *Proc. 12th International Conference on Learning Representations* https://openreview.net/pdf?id=CfXh93NDgH (ICLR,2024).
43. Talat, Z. et al. You reap what you sow: on the challenges of bias evaluation under multilingual settings. In *Proc. BigScience Episode #5–Workshop on Challenges & Perspectives in Creating Large Language Models* (eds Fan, A. et al.) 26–41 (Association for Computational Linguistics, 2022).
44. Kreutzer, J. et al. Quality at a glance: an audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics* **10**, 50–72 (2022).
45. Shankar, S. et al. No classification without representation: assessing geodiversity issues in open data sets for the developing world. Preprint at https://arxiv.org/abs/1711.08536 (2017).
46. De Vries, T., Misra, I., Wang, C. & Van der Maaten, L. Does object recognition work for everyone? In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 52–59 (IEEE, 2019).
47. Mahadev, R. & Chakravarti, A. Understanding gender and racial disparities in image recognition models. Preprint at https://arxiv.org/abs/2107.09211 (2021).
48. Ahia, O., Kreutzer, J. & Hooker, S. The low-resource double bind: an empirical study of pruning for low-resource machine translation. In *Proc. Findings of the Association for Computational Linguistics: EMNLP 2021* (eds Moens, F.-M. et al.) 3316–3333 (ACM, 2021).
49. Epstein, Z. et al. Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
50. Quang, J. Does training AI violate copyright law? *Berkeley Technol. L. J.* **36**, 1407 (2021).
51. Lee, K., Cooper, A. F. & Grimmelmann, J. Talkin 'bout AI generation: copyright and the generative-AI supply chain. *J. Copyright Soc. USA* (in the press).
52. Gervais, D. J. AI derivatives: the application to the derivative work right to literary and artistic productions of AI machines. *Seton Hall Law Rev.* **52**, 1111 (2021).
53. Henderson, P. et al. Foundation models and fair use. *J. Mach. Learn. Res.* **24**, 1–79 (2023).
54. Lemley, M. A. & Casey, B. Fair learning. *Texas L. Rev.* **99**, 743 (2020).
55. Sobel, B. L. W. Artificial intelligence's fair use crisis. *Columbia J. L. Arts* **41**, 45–97 (2017).
56. Samuelson, P. Generative AI meets copyright. *Science* **381**, 158–161 (2023).
57. Doe v. GitHub, Inc., 22-cv-06823-JST (N.D. Cal. 2024).
58. Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605 (2d Cir. 2006).
59. Grossman, C. A. From Sony to Grokster, the failure of the copyright doctrines of contributory infringement and vicarious liability to resolve the war between content and destructive technologies. *Buffalo L. Rev.* **53**, 141–268 (2005).
60. Marks, C. P. & Moll, D. K. *The Law of Business Torts and Unfair Competition: Cases, Materials, and Problems*. American Casebook Series (West Academic, 2023).
61. Victor, J. & Efrati, A. Alphabet's Google and DeepMind pause grudges, join forces to chase OpenAI. *The Information* https://www.theinformation.com/articles/alphabets-google-and-deepmind-pause-grudges-join-forces-to-chase-openai (2023).
62. Suggs, N. & Venables, P. Protecting customers with generative AI indemnification. *Google Cloud* https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification (2023).
63. Mahari, R. et al. Comment to U.S. copyright office on data provenance and copyright (US Copyright Office, 2023); https://dspace.mit.edu/handle/1721.1/154171
64. Longpre, S. et al. Position: data authenticity, consent, & provenance for AI are all broken: what will it take to fix them? *An MIT Exploration of Generative AI* https://doi.org/10.21428/e4baedd9.a650f77d (2024).
65. Kinney, R. M. et al. The semantic scholar open data platform. Preprint at https://arxiv.org/abs/2301.10140 (2023).
66. Petrov, A., La Malfa, E., Torr, P. & Bibi, A. Language model tokenizers introduce unfairness between languages. In *Proc. of the 37th International Conference on Neural Information Processing Systems* 36963–36990 (Curran, 2024).
67. Bender, E. M. & Friedman, B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics* **6**, 587–604 (2018).

68. Pushkarna, M., Zaldivar, A. & Kjartansson, O. Data cards: purposeful and transparent dataset documentation for responsible AI. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 1776–1826 (ACM, 2022).
69. Bender, E. M. On achieving and evaluating language-independence in NLP. *Linguist. Issues Lang. Technol.* https://doi.org/10.33011/lilt.v6i.1239 (2011).
70. Longpre, S. et al. Data-Provenance-Initiative/Data-Provenance-Collection: Data Provenance Initiative Release. *Zenodo* https://doi.org/10.5281/zenodo.11587503 (2024).
71. Durbin, J. Airoboros: using large language models to fine-tune large language models. *GitHub* https://github.com/jondurbin/airoboros (2023).
72. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint at https://arxiv.org/abs/2204.05862 (2022).
73. Ganguli, D. et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. Preprint at https://arxiv.org/abs/2209.07858 (2022).
74. Xu, C., Guo, D., Duan, N. & McAuley, J. Baize: an open-source chat model with parameter-efficient tuning on self-chat data. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 6268–6278 (Association for Computational Linguistics, 2023).
75. Kryściński, W., Rajani, N., Agarwal, D., Xiong, C. & Radev, D. Booksum: a collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (eds Goldberg, Y. et al.) 6536–6558 (Association for Computational Linguistics, 2022).
76. Li, G., Hammoud, H., Itani, H., Khizbullin, D. & Ghanem, B. CAMEL: communicative agents for 'mind' exploration of large scale language model society. In *Proc. of the 37th International Conference on Neural Information Processing Systems* 51991–52008 (Curran, 2024).
77. Kim, S. et al. The CoT collection: improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 12685–12708 (Association for Computational Linguistics, 2023).
78. Muennighoff, N. et al. Octopack: instruction tuning code large language models. In *Proc. 12th International Conference on Learning Representations* https://openreview.net/pdf?id=mw1PWNSWZP (ICLR, 2024).
79. Conover, M. et al. Free Dolly: introducing the world's first truly open instruction-tuned LLM. *Databricks* www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm (2023).
80. Peng, B., Li, C., He, P., Galley, M. & Gao, J. Instruction tuning with GPT-4. Preprint at https://arxiv.org/abs/2304.03277 (2023).
81. Anand, Y., Nussbaum, Z., Duderstadt, B., Schmidt, B. & Mulyar, A. GPT4all: training an assistant-style chatbot with large scale data distillation from GPT-3.5-turbo. GitHub https://github.com/nomic-ai/gpt4all (2023).
82. Patil, S. G., Zhang, T., Wang, X. & Gonzalez, J. E. Gorilla: large language model connected with massive APIs. Preprint at *arXiv* https://arxiv.org/abs/2305.15334 (2023).
83. Guo, B. et al. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. Preprint at https://arxiv.org/abs/2301.07597 (2023).
84. Nguyen, H., Suri, S., Tsui, K. & Schuhmann, C. *The Open Instruction Generalist (OIG) Dataset* (LAION, 2023); https://laion.ai/blog/oig-dataset/
85. Zhou, C. et al. Lima: Less is more for alignment. In *Proc. of the 37th International Conference on Neural Information Processing Systems* 55006–55021 (Curran, 2024).
86. Köksal, A., Schick, T., Korhonen, A. & Schütze, H. Longform: optimizing instruction tuning for long text generation with corpus extraction. Preprint at https://arxiv.org/abs/2304.08460 (2023).
87. Stiennon, N. et al. Learning to summarize from human feedback. In *Proc. of the 34th International Conference on Neural Information Processing Systems* 3008–3021 (Curran, 2020).
88. Köpf, A. et al. OpenAssistant conversations–democratizing large language model alignment. In *Proc. of the 37th International Conference on Neural Information Processing Systems* 47669–47681 (Curran, 2024).
89. Mukherjee, S. et al. Orca: progressive learning from complex explanation traces of GPT-4. Preprint at https://arxiv.org/abs/2306.02707 (2023).
90. Ethayarajh, K., Zhang, H., Wang, Y. & Jurafsky, D. *Stanford Guman Preferences Dataset* (2023); https://huggingface.co/datasets/stanfordnlp/SHP
91. Vercel. *Sharegpt* https://sharegpt.com/ (2023).
92. Li, R. et al. Starcoder: may the source be with you! *Trans. Mach. Learn. Res.* https://openreview.net/pdf?id=KoFOg41haE (2023).
93. Sileo, D. tasksource: a dataset harmonization framework for streamlined NLP multi-task learning and evaluation. Preprint at https://arxiv.org/abs/2301.05948 (2023).
94. Weston, J. et al. Towards AI-complete question answering: a set of prerequisite toy tasks. In *Proc. of the 4th International Conference on Learning Representations* (eds Bengio, Y. & and LeCun, Y.) (ICLR, 2016).
95. Eldan, R. & Li, Y. Tinystories: how small can language models be and still speak coherent english? Preprint at https://arxiv.org/abs/2305.07759 (2023).
96. Qin, Y. et al. ToolLLM: facilitating large language models to master 16000+ real-world APIs. In *Proc. 2024 International Conference on Learning Representations* https://openreview.net/pdf?id=dHng2O0Jjr (ICLR, 2024).
97. Ding, N. et al. Enhancing chat language models by scaling high-quality instructional conversations. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 3029–3051 (Association for Computational Linguistics, 2023).
98. Honovich, O., Scialom, T., Levy, O. & Schick, T. Unnatural instructions: tuning language models with (almost) no human labor. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Rogers, A. et al.) 14409–14428 (Association for Computational Linguistics, 2023).
99. Nakano, R. et al. WebGPT: browser-assisted question-answering with human feedback. Preprint at https://arxiv.org/abs/2112.09332 (2021).
100. Hendrycks, D. et al. Measuring massive multitask language understanding. In *Proc. International Conference on Learning Representations* (2020).
101. Srivastava, A. et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.* https://openreview.net/pdf?id=uyTL5Bvosj (2023).

Acknowledgements

We thank K. Lee, A. F. Cooper, P. Henderson, A. Skowron and S. Biderman for valuable comments and feedback.

Author contributions

We emphasize that all authors contributed crucial elements to this project, and core contributors in particular are recognized with hands

on service to the design and construction of Data Provenance's first implementation. S.L. was primary designer and coder of the repository and explorer interface, and led audit implementation and analysis, as well as the manual annotation process. R.M. led the legal analysis and licensing annotation design. A.C. led automatic inferencing of dataset text metrics, topics and task category annotations, and supported writing, analysis and code testing. N.O.-M. led visualization design, particularly interactive visualizations in the DPExplorer. D.S. led data aggregator linking and metadata crawling, and supported writing, analysis, source annotation and adding datasets. W.B. added eight data collections and supported writing and data analysis. N.M. added several large data collections and supported writing, analysis, visualization and source annotations. N.K. led licensing annotation effort and supported adding datasets along with testing. J.K. was an advisor and led the text source annotation effort and supported with framing, writing and analysis. K.P. added several datasets and supported writing, analysis and dataset preparation for Hugging Face. X.(A.)W. added several datasets, did testing and supported automatic metadata collection. E.S. led final dataset preparation for Hugging Face upload and testing. K.B. was an advisor on project design and framing. T.W. was an advisor, particularly on data analysis and visualizations, and supported writing and DPExplorer design. L.V. was an advisor on data copyright and licensing, and supported writing in the legal discussion section. S.P. was an advisor on general project design and framing. S.H. was an advisor on general project design and framing, as well as supporting writing, analysis and directing experiments.

Competing interests

The following authors are employed by a firm engaged in AI or related research: N.M. is a Research Engineer at Contextual AI. K.P. is a Research Scientist at Apple. E.S. is CEO of Teraflop AI. K.B. is Director of Engineering at MLCommons. L.V. is cofounder and general counsel of Tidelift. S.H. is head of Cohere For AI. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00878-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00878-8.

Correspondence and requests for materials should be addressed to Robert Mahari.

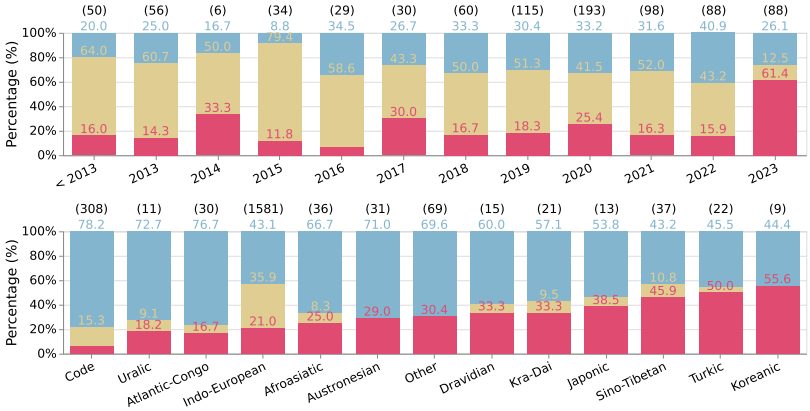
Peer review information *Nature Machine Intelligence* thanks Thomas Burri and Nick Vincent for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

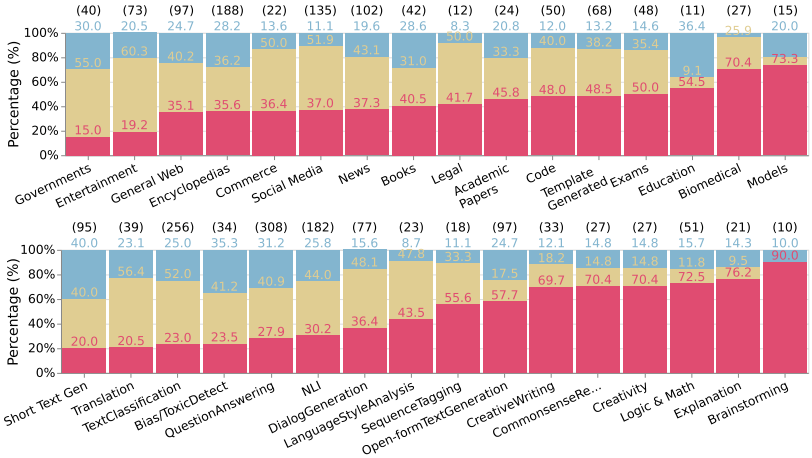
[1]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Harvard Law School, Harvard University, Cambridge, MA, USA. [3]Department of Computer Science, University of California, Irvine, CA, USA. [4]Center for Constructive Communication, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Inria Centre, University of Lille, Lille, France. [6]Contextual AI, Mountain View, CA, USA. [7]College of Engineering & Applied Science, University of Colorado Boulder, Boulder, CO, USA. [8]Data Provenance Initiative, Cambridge, MA, USA. [9]Olin College of Engineering, Needham, MA, USA. [10]Teraflop AI, Boca Raton, FL, USA. [11]ML Commons, San Francisco, CA, USA. [12]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA. [13]Tidelift, Boston, MA, USA. [14]Cohere For AI, Toronto, Ontario, Canada. [15]These authors contributed equally: Shayne Longpre, Robert Mahari. ✉e-mail: rmahari@mit.edu

https://doi.org/10.1038/s42256-024-00878-8



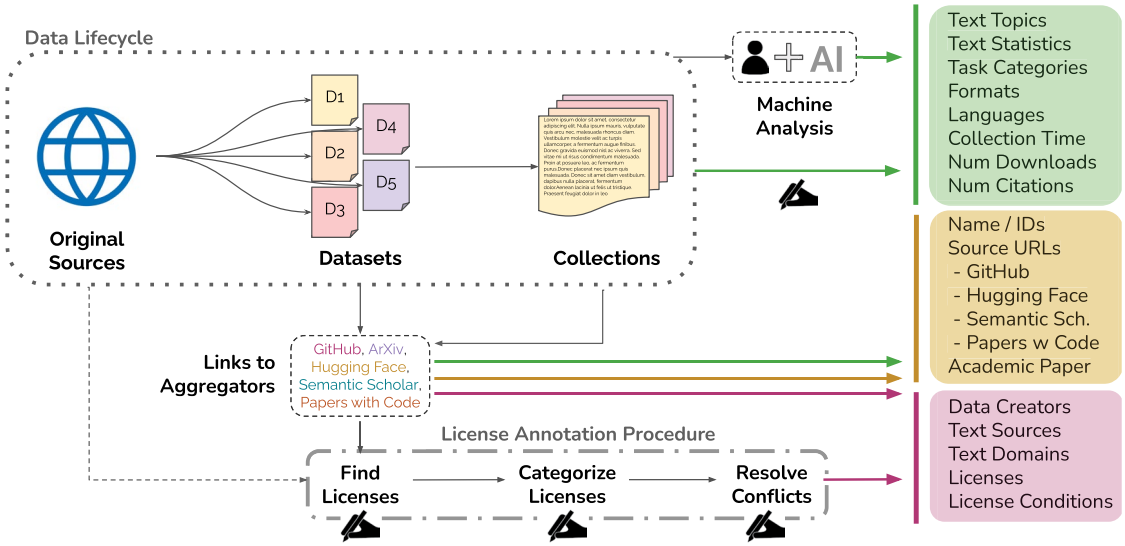
Extended Data Fig. 1 | Licenses over time and across languages. The distribution of datasets in each time of collection (top) and language family (bottom) category, with total count above the bars, and the portion in each license use category shown via bar colour. Red represents Non-commercial/Academic-Only, Yellow represents Unspecified, and Blue represents Commercial. Lower-resource languages, and datasets created in 2023 see a spike in non-commercial licensing.

https://doi.org/10.1038/s42256-024-00878-8



Extended Data Fig. 2 | Licenses across domain sources and tasks. 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 colour. Red represents Non-commercial/Academic-Only, Yellow represents Unspecified, and Blue represents 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.

https://doi.org/10.1038/s42256-024-00878-8



Extended Data Fig. 3 | DPCollection annotation pipeline. The annotation pipeline uses human and human-assisted procedures to annotate dataset Identifiers, Characteristics, and Provenance. The Data Lifecycle is traced, from the original sources (web crawls, human or synthetic text), to curated datasets and packaged collections. Information is collected at each stage, not just the last. The License Annotation Procedure is described in the section on license collection.

https://doi.org/10.1038/s42256-024-00878-8

Extended Data Table 1 | Licenses and citations for the dataset collections presented in this paper

	Cite	Licenses
Collection		
Airoboros	71	CC BY-NC 4.0
Alpaca	2	CC BY-NC 4.0
Anthropic HH	72,73	MIT License
BaizeChat	74	CC BY-NC 4.0
BookSum	75	Academic Only
CamelAI Sci.	76	CC BY-NC 4.0
CoT Coll.	77	Non Commercial
Code Alpaca	–	Unspecified
CommitPackFT	78	Various
Dolly 15k	79	CC BY-SA 3.0
Evol-Instr.	42	Academic Only
Flan Collection	17	Various
GPT-4-Alpaca	80	CC BY-NC 4.0
GPT4All	81	Various
GPTeacher	–	Unspecified
Gorilla	82	Apache License 2.0
HC3	83	Various
Joke Expl.	–	MIT License
LAION OIG	84	Various
LIMA	85	CC BY-NC-SA 4.0
Longform	86	CC BY-SA 3.0, Unspecified, CC BY-SA 4.0
OpAsst OctoPack	78	CC BY 4.0
OpenAI Summ.	87	CC BY 4.0
OpenAssistant	88	CC BY 4.0
OpenOrca	89	Various
SHP	90	Unspecified
Self-Instruct	5	Apache License 2.0
ShareGPT	91	Unspecified
StackExchange	–	Unspecified
StarCoder	92	BigScience OpenRAIL-M
Tasksource Ins.	93	Various
Tasksource ST	94	Various
TinyStories	95	CDLA Sharing 1.0
Tool-Llama	96	CC BY-NC 4.0
UltraChat	97	CC BY-NC 4.0
Unnatural Instr.	98	MIT License
WebGPT	99	Apache License 2.0, CC BY-SA 4.0
xP3x	12	Various

Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. More comprehensive details are available at https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection/tree/main/data_summaries. Note that we remove datasets related to common benchmarks like MMLU[100] and BigBench[101].

https://doi.org/10.1038/s42256-024-00878-8

Extended Data Table 2 | Summary of Creators, Topics, and Source Domains for all data. A summary of the distribution of Creators, Topics, and Source Domains across all 1800+ datasets. Datasets can have multiple creators, text topics, and sources

Name	Pct
Academic	68.7%
University of Washington	8.9%
Stanford University	6.8%
New York University	5.4%
University of Southern...	3.5%
Carnegie Mellon Univer...	3.5%
Saarland University	2.6%
Cardiff University	2.3%
Industry Lab	21.4%
Facebook AI Research	8.4%
Microsoft Research	4.1%
Google Research	2.9%
DeepMind	1.9%
Microsoft Semantic Mac...	0.9%
NAVER AI Lab	0.8%
Salesforce Research	0.7%
Research Group	17.1%
AI2	12.3%
CLUE team	0.5%
Alan Turing Institute	0.5%
CodeX	0.4%
Qatar Computing Resear...	0.4%
Barcelona Supercomputi...	0.4%
BigCode	0.2%
Corporation	15.8%
Google	2.1%
IBM	2.0%
Microsoft	1.4%
Wind Information Co.	1.4%
Snap Inc.	1.3%
Meta	1.1%
Synapse Développement	1.1%
Startup	4.0%
OpenAI	1.3%
NomicAI	0.8%
Omniscien Technologies	0.4%
Anthropic AI	0.2%
EightSleep	0.2%
Curai	0.2%
IMRSV Data Labs	0.2%
Other	0.7%

(a) **Creators**

Name	Pct
Question Answering	36.0%
Question Answering	27.7%
Multiple Choice Questi...	3.9%
Information Extraction	1.8%
Text Classification	29.9%
Text Classification	16.1%
Sentiment Analysis	9.8%
Named Entity Recognition	4.3%
NLI	21.1%
Textual Entailment	14.6%
Natural Language Inference	5.3%
Fact Verification	1.3%
Open-form Text Gen...	11.3%
Open-form Text Generation	2.2%
Title Generation	1.5%
Inverted Summarization	1.2%
Short Text Generation	10.9%
Question Generation	4.0%
Fill in The Blank	1.4%
Inverted Multiple-Choice	0.9%
Dialog Generation	9.0%
Dialogue Generation	4.2%
Dialog Generation	3.7%
Dialogue Act Recognition	0.4%
Summarization	6.3%
Summarization	5.7%
Simplification	0.5%
Summarization of US Co...	0.1%
Logical and Math Reason...	6.0%
Logical Reasoning	2.3%
Data Analysis	2.0%
Algebraic Expression E...	1.2%
Code	4.8%
Response Ranking	4.4%
Translation	4.4%
Creative Writing	3.9%
Other	23.9%

(b) **Topics**

Name	Pct
Encyclopedias	21.5%
wikipedia.org	14.6%
wikihow.com	2.7%
dbpedia	1.4%
Social Media	15.9%
reddit	6.2%
twitter	4.0%
quora	1.6%
General Web	11.2%
undisclosed web	7.0%
commoncrawl.org	2.5%
data.world/samayo/coun...	0.6%
News	11.1%
cnn.com	1.6%
financial news	1.5%
press releases	1.4%
Entertainment	8.5%
opensubtitles.org	2.5%
imdb.com	1.6%
travel guides	1.3%
Code	5.7%
stackexchange.com	2.0%
github	1.2%
opus software projects	0.9%
Exams	5.6%
web exams	2.9%
gmat	1.1%
gre exams	0.9%
Books	4.9%
project gutenberg	2.0%
non-fiction books	1.3%
fiction books	1.3%
Governments	4.7%
Biomedical	3.2%
Search Queries	3.0%
Academic Papers	2.8%
Other	61.2%

(c) **Domains & Sources**

nature portfolio

Corresponding author(s): Robert Mahari

Last updated by author(s): 6/14/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	This data was collected with a mix of manual and automated techniques, leveraging dataset aggregators, namely GitHub, Hugging Face and Semantic Scholar. Annotating and verifying license information, in particular, required a carefully guided manual workflow, designed with legal practitioners. Once these information aggregators were connected, it was possible to synthesize or scrape additional metadata, such as dataset languages, task categories, and time of collection. And for richer details on each dataset, like text topics and source, we used carefully tuned prompts on language models inspecting each dataset. The sources we analyze are listed on our public repository: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection/tree/main/data_summaries.
Data analysis	Our data analysis and collection pipeline included both manual collection and automated data preparation/analysis using latest standard libraries at the time of submission. Specifically, we use the following Python (Python version 3.8.9) packages: aiohttp (version 3.9.5), aiosignal (version 1.3.1), annotated-types (version 0.7.0), anyio (version 4.4.0), async-timeout (version 4.0.3), attrs (version 23.2.0), certifi (version 2023.7.22), chardet (version 5.2.0), charset-normalizer (version 3.3.2), ConfigArgParse (version 1.7), datasets (version 2.19.2), dill (version 0.3.8), distlib (version 0.3.6), distro (version 1.9.0), exceptiongroup (version 1.2.1), filelock (version 3.11.0), frozenlist (version 1.4.1), fsspec (version 2024.3.1), h11 (version 0.14.0), httpcore (version 1.0.5), httpx (version 0.27.0), huggingface-hub (version 0.23.3), idna (version 3.4), jsonlines (version 4.0.0), multidict (version 6.0.5), multiprocess (version 0.70.16), numpy (version 1.24.4), openai (version 1.33.0), packaging (version 24.1), pandas (version 2.0.3), platformdirs (version 3.2.0), pyarrow (version 16.1.0), pyarrow-hotfix (version 0.6), pydantic (version 2.7.3), pydantic_core (version 2.18.4), python-dateutil (version 2.9.0.post0), python-dotenv (version 1.0.1), pytz (version 2024.1), PyYAML (version 6.0.1), requests (version 2.32.3), semanticscholar (version 0.5.0), sniffio (version 1.3.1), tabulate (version 0.9.0), tenacity (version 8.2.3), tqdm (version 4.66.4), typing_extensions (version 4.12.2), tzdata (version 2024.1), urllib3 (version 2.1.0), virtualenv (version 20.21.0), xxhash (version 3.4.1), yarl (version 1.9.4).

All details and code are documents in our public repo: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used in our analysis, including the manually collected data, as well as a generalizable pipeline for future data collection, can be found in our public repository: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection. A full list of data sources may be found at: https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection/tree/main/data_summaries. These repositories contain all of the metadata we collected, and build out downloaders that pull from Hugging Face or GitHub, to standardize formats, wrap them in their metadata, and then apply tools to filter, sort, select, and visualize those datasets.

From these collections, we identify text datasets for multi-task finetuning, preference/human feedback tuning, and multi-turn dialog. These are selected by compiling popular datasets on Hugging Face, for a diverse set of tasks, as well as other popular datasets we discovered in the process of investigating popular instruction tuned models on Hugging Face for general-purpose chatting, tool-use, multilingual Q&A, and other common NLP tasks. Although this process is partly subjective, we devise an annotation pipeline to maximize reproducibility. The annotated data may be accessed, visualized, and explored on https://dataprovenance.org/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We convene a multi-disciplinary effort between legal and machine learning experts to systematically audit and trace 1800+ text datasets. We develop tools and standards to trace the lineage of these datasets, from their source, creators, series of license conditions, properties, and subsequent use (this includes both qualitative and quantitative data).
Research sample	1,800 text finetuning datasets included in the most widely used 44 data collections. While this sample is not fully representative of all text finetuning datasets, it was compiled based on Hugging Face usage statistics to ensure that we focus on the most widely used datasets. The data on which we base our research comes from manual and automated annotation of data hosting sites (Hugging Face, Papers with Code, GitHub) and the dataset papers.
Sampling strategy	The research sample consisted of the most widely used data collections on Hugging Face, which see 100s to 10M+ monthly downloads. The goal of our study is not to provide a total audit of text fine tuning data but to trace the provenance of the most widely used data and thus our focus on download metrics ensures a sample that includes the most commonly used data.

Data collection	Data was automatically scraped and manually labeled as appropriate. While researchers were not blinded to study hypotheses, the data collection procedure (described in our methods) was independent of these hypotheses.
Timing	May - September 2023
Data exclusions	No exclusions
Non-participation	No participants were involved in the study.
Randomization	The study sample was based on the most frequently used datasets according to Hugging Face downloads and therefore randomization is not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging