



HAL
open science

The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & 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.

► To cite this version:

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, et al.. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. 2023. hal-04290233

HAL Id: hal-04290233

<https://hal.science/hal-04290233v1>

Preprint submitted on 16 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI

Shayne Longpre^{1†} Robert Mahari^{1,2} Anthony Chen³ Naana Obeng-Marnu^{1,4}
Damien Sileo⁵ William Brannon^{1,4} Niklas Muennighoff⁶ Nathan Khazam⁷
Jad Kabbara^{1,4} Kartik Perisetla Xinyi (Alexis) Wu⁸ Enrico Shippole Kurt Bollacker⁷
Tongshuang Wu⁹ Luis Villa¹⁰ Sandy Pentland¹ Sara Hooker¹¹

¹ MIT ² Harvard Law School ³ UC Irvine ⁴ MIT Center for Constructive Communication
⁵ Inria, Univ. Lille Center ⁶ Contextual AI ⁷ ML Commons ⁸ Olin College
⁹ Carnegie Mellon University ¹⁰ Tidelifit ¹¹ Cohere For AI

Abstract

The race to train language models on vast, diverse, and inconsistently documented datasets has raised pressing concerns about the legal and ethical risks for practitioners. To remedy these practices threatening data transparency and understanding, 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. Our landscape analysis highlights the sharp divides in composition and focus of commercially open vs closed datasets, with closed datasets monopolizing important categories: lower resource languages, more creative tasks, richer topic variety, newer and more synthetic training data. This points to a deepening divide in the types of data that are made available under different license conditions, and heightened implications for jurisdictional legal interpretations of copyright and fair use. We also observe frequent miscategorization of licenses on widely used dataset hosting sites, with license omission of 70%+ and error rates of 50%+. This points to a crisis in misattribution and informed use of the most popular datasets driving many recent breakthroughs. As a contribution to ongoing improvements in dataset transparency and responsible use, we release our entire audit, with an interactive UI, the **Data Provenance Explorer**, which allows practitioners to trace and filter on data provenance for the most popular open source finetuning data collections: www.dataprovenance.org.

1 Introduction

The latest wave of language models, both public (Chung et al., 2022; Taori et al., 2023; Geng et al., 2023) and proprietary (Anil et al., 2023; OpenAI, 2023; Anthropic, 2023; Yoo et al., 2022) attribute their powerful abilities in large part to the diversity and richness of ever larger training datasets, including pre-training corpora, and finetuning datasets compiled by academics (Wei et al., 2021; Sanh et al., 2021; Muennighoff et al., 2022), synthetically generated by models (Taori et al., 2023; Wang et al., 2022a), or aggregated by platforms like Hugging Face (Lhoest et al., 2021). Recent trends see practitioners combining and re-packaging thousands of datasets and web sources (Gao et al., 2020; Penedo et al., 2023; Wang et al., 2022b; Longpre et al., 2023a), but despite some notable documentation efforts (Spacerini, 2021; Biderman et al., 2022), there are diminishing efforts to attribute, document or understand the raw ingredients into new models (Dodge et al., 2021; Bandy and Vincent, 2021; Bommasani et al., 2023a).

[†] Correspondence: data.provenance.init@gmail.com

A Crisis in Data Transparency & its Consequences. Increasingly, widely used dataset collections are treated as monolithic, instead of a lineage of data sources, scraped (or model generated), curated, and annotated, often with multiple rounds of re-packaging (and re-licensing) 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 the increased copyright scrutiny (Saveri et al., 2023). Together, these factors have seen fewer Datasheets (Gebru et al., 2021), non-disclosure of training sources (OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023), and ultimately a decline in understanding training data (Sambasivan et al., 2021b; Longpre et al., 2023b).

This lack of understanding can lead to data leakages between training and test data (Elangovan et al., 2021; Carlini et al., 2022), expose personally identifiable information (PII) (Bubeck et al., 2023), present unintended biases or behaviours (Welbl et al., 2021; Xu et al., 2021; Pozzobon et al., 2023), 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 (e.g., WizardCoder (Luo et al., 2023) licensed for commercial use, while training on commercially-prohibited OpenAI data), license revisions post-public release (with MPT-StoryTeller (Frankle, 2023)), and even copyright lawsuits (e.g. Stability AI (Arstechnica, 2023) and OpenAI (Saveri et al., 2023)). 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 datasets’ sourcing, creation and licensing heritage, as well as its characteristics, *Data Provenance*.

Unreliable Data Provenance & 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 44 of the most widely used and adopted text data collections, spanning 1800+ finetuning datasets. We compile and expand relevant metadata with a much richer taxonomy than Hugging Face, Papers with Code, or other aggregators (see Section 2.1). With legal experts, we design a pipeline for tracing dataset provenance, including the original source of the dataset, the associated licenses, creators, and subsequent use.

As a byproduct of our work establishing the *Data Provenance* of widely used datasets, we are able to characterize the AI data ecosystem/supply chain (Cen et al., 2023; Bommasani et al., 2023c), as well as state of the field for policymakers, researchers and legal experts. Our work points to a crisis in license laundering and informed usage of popular datasets, with systemic problems in sparse, ambiguous, or incorrect license documentation. Notably, we find that 70%+ of licenses 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. Second, the licenses that are attached to datasets uploaded to dataset sharing platforms are often inconsistent with the license ascribed by the original author of the dataset—our rigorous re-annotation of licenses finds that 66% of analyzed Hugging Face licenses were in a different use category, often labeled as more permissive than the author’s intended license. As a result, much of this data is risky to use (or harmfully misleading) for practitioners who want to respect the data provenance of a work. Our initiative reduces “Unspecified” licenses from 72%+ to 30% and attaches license URLs for under-resourced 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 1800+ text datasets (the “*DPCollection*”), their licenses, conditions, and sources. We demonstrate a growing adoption and reliance on software licenses in the AI community and synthesize observations into legal guidance for developers (Section 4).
2. **The Data Provenance Explorer (DPEXplorer)***, 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.

*www.dataprovenance.org

2 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 Section 2.1). As a first contribution of the initiative, we audit 44 instruction or “alignment” finetuning data collections composed of 1858 individual datasets, selected by experts for their widespread adoption and use in the community. The selected collections and their variants see 100s to 10M+ monthly downloads on Hugging Face, with the datasets within these collections tallying to many more Table 1.

The 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 (Ouyang et al., 2022). 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. We remove datasets related to common benchmarks like MMLU (Hendrycks et al., 2020) and BigBench (Srivastava et al., 2023).

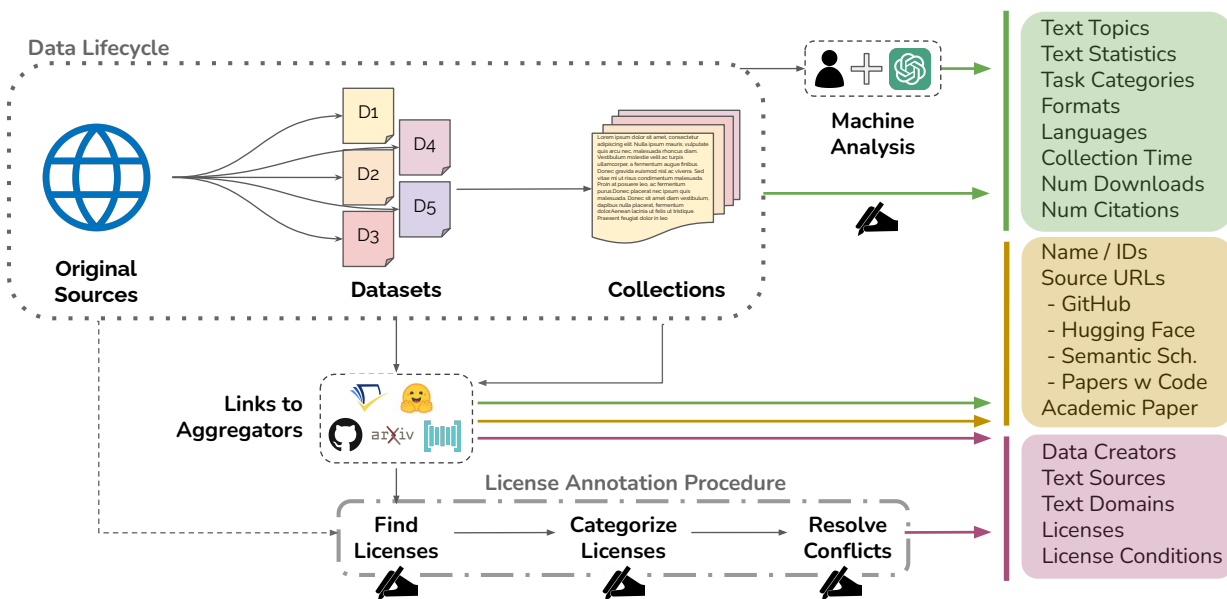


Figure 1: The DPCollection annotation pipeline uses human and human-assisted procedures to annotate dataset **Identifiers**, **Characteristics**, and **Provenance**. The *Data Lifecycle* is traced, from the original sources (web scrapes, human or synthetic text), to curated datasets and packaged collections. Information is collected at each stage, not just the last. The *License Annotation Procedure* is described in Section 2.2.

2.1 Data Provenance Explorer (DPExplorer)

Our information audit spans (I) *identifier information*, bridging metadata from several aggregators, including Hugging Face, GitHub, Papers with Code, Semantic Scholar, and ArXiv, (II) detailed *dataset characteristics* for a richer understanding of training set composition, and (III) *dataset provenance* for licensing and attribution. We expand our provenance metadata beyond just licenses, because conversations with practitioners revealed they rely not only on data licenses, but on a specific *legal & ethical risk tolerance*, parameterized by (a) the lineage of licenses, (b) the data source, (c) the creator’s identity, and (d) the precedence of adoption by other developers.

We release our extensive audit, as two tools: (1) a data explorer interface, the *Data Provenance Explorer (DPExplorer)* for widespread use, and (2) an accompanying repository for practitioners to download the data

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

Table 1: **Alignment tuning collections and their characteristics.** Properties of the collections include the numbers of datasets, dialogs, unique tasks, languages, topics, text domains, Huggingface 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 (🌐), or model generated text (🤖). The dialog formats of each collection can be: zero-shot (Z), few-shot (F), chain-of-thought (C), response ranking (R), and multi-turn dialog (M). The USE column indicates whether a collection includes data licensed for commercial use (●), data with no license (“unspecified”: ●), data only licensed for non-commercial or academic use (●). *Note that these licenses are self-reported and their applicability is complicated, requiring legal consultation.* The “O” column indicates if the collection includes OpenAI model generations, which may or may not affect commercial viability (see Section 4)

filtered for license 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 licenses (Section 2.3). Modern researchers training on hundreds of datasets often find it onerous to manually curate extensive data cards for these compilations (Mitchell et al., 2019; Gebru et al., 2021). 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.

Collecting comprehensive metadata for each dataset required leveraging several sources including collection by linking to resources already on the web (🌐), human annotation by legal experts (⚖️), or using GPT-4 to assist in human annotation (🤖).

Identifier Information discloses links and connects aggregator identifiers.

1. **Dataset Identifiers** 📄: The dataset’s name, associated paper title, and description of the dataset.
2. **Dataset Aggregator Links** 🔗: 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** 📁: The name and URL to the data collection of which this dataset is a part.

Dataset Characteristics detail information relevant to understanding data representation/composition, and curating a training set.

1. **Languages** 🗣️: Each of the languages represented in the dataset, so developers can easily follow the “Bender Rule” (Bender, 2011).
2. **Task Categories** 📁🤖: The 20+ task categories represented in the instructions, such as Question Answering, Translation, Program Synthesis, Toxicity Identification, Creative Writing, and Roleplaying.
3. **Text Topics** 🗣️: An automated annotation of the topics discussed in the datasets, with GPT-4 labeling a sample of 100 examples for up to 10 covered topics.
4. **Text Length Metrics** 📏: The minimum, maximum, and mean number of dialog turns per conversation, of characters (agnostic to tokenization/non-whitespace languages, as this introduces biases (Petrov et al., 2023)) per user instruction and assistant responses.
5. **Format** 📄: The format and intended use of the data. The options are zero-shot prompts, few-shot prompts, chain-of-thought prompts, multi-turn dialog, and response ranking.
6. **Time of Collection** 🌐: The time as which the work was published, which acts as an upper bound estimate of the age of the text.

Dataset Provenance

1. **Licenses** 🌐📄: The license name and URLs associated with the data, using the process described in Section 2.2. We also enable filtering by license use classes, categorized by legal professionals.
2. **Text Source** 📄🗣️: The original sources of the text, often Wikipedia, Reddit, or other scraped online/off-line sources.
3. **Creators** 📄: The institutions of the dataset authors, including universities, corporations, and other organizations.
4. **Attribution** 🌐: The attribution information for the authors of the paper associated with the dataset.
5. **Citation & Download Counts** 🌐: 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 .

2.2 License Annotation Process

One of our central contributions is to validate the licenses associated with widely used and adopted datasets. This followed a time-intensive human annotation protocol, to collect dataset authors’ self-reported licenses, and categorize them according to stated conditions. Note that this protocol reflects best efforts to verify self-reported licenses, and does not constitute legal advice (see Section 4). Additionally, it is important to note that the enforceability of these licenses depends on several factors discussed in Section 4. 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 this 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 license annotation workflow follows these steps:

1. **Compile all Self-Reported License Information** We aggregate all licensing information reported on GitHub, ArXiv, Hugging Face, Papers with Code, and the collection itself (e.g. Super-Natural Instructions, Wang et al. (2022c)).
2. **Search for explicit Data Licenses** The annotator searches for a license specifically given to the dataset (*not the accompanying code*) by the authors. A license is found if (a) the GitHub repository mentions or links a license in reference to the data, (b) the Hugging Face license label was uploaded by the dataset creator themselves, (c) the paper, Hugging Face, or Papers with Code provide a dataset-specific license link, attributable to the data authors.
3. **Identify a License Type** A license may fall into a set of common types (e.g. MIT, Apache 2, CC BY SA, etc.), be a “Custom” license, a permission Request Form, or if none was found for the data, *Unspecified*. If a dataset has multiple licenses, the annotator will list each of them, according to their types.
4. **Categorize Licenses** From the perspective of a machine learning practitioner, licensing typically is viewed through the lens of how it impacts the model lifecycle—does it impede or allow for training on the data, downstream use conditions, attributing, modifying or re-distributing it. Based on discussions with industry experts, we categorize licenses based on three important features that impact 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 licensed under the same terms as the original (**Share-Alike**). If there are multiple licenses for a dataset, its categorization for each feature is the chosen as the strictest across licenses.
5. **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 licenses, even when it is ambiguous whether these apply to the code or the data. They may also wish to include or exclude datasets based on whether these are already widely used in practice, where the original data was sourced from, and if the creator is a competitor. To supplement the above license categories, we also collect all this metadata for fine-grained selection and filtering.

2.3 Data Provenance Card—A Data Bibliography

Prior work has stressed the importance of data documentation and attribution (Bender and Friedman, 2018; Bommasani et al., 2023a). In particular, Gebru et al. (2021)’s Datasheets breaks down documentation into motivation, composition, collection process, processing, uses, maintenance, and distribution. Similarly, Bender and Friedman (2018) 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.[†]

[†]Auto-generated at <https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection>

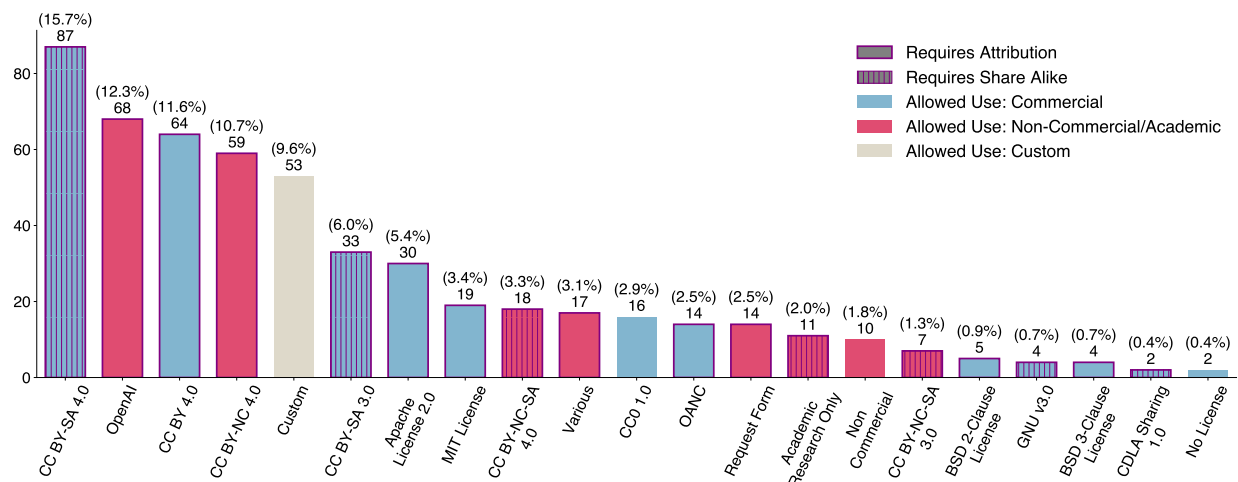


Figure 2: We plot the distributions of licenses used in the DPCollection, a popular sample of the major supervised NLP datasets. We find a long tail of custom licenses, adopted from software for data. 73% of all licenses require attribution, and 33% share-alike, but the most popular are usually commercially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

Our collected schema allows this store to succinctly capture the attribution (links to repositories, aggregator copies, papers, creators), provenance (text/machine sources, licenses), 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 prior work, while providing some advantages from its structure. First, the Data Provenance Card can be easily searched, sorted, filtered and analyzed, 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 re-packaged and absorbed into newer and bigger collections, Data Provenance Cards are easily adaptable by simply appending or concatenating them together. Altogether, we hope this tooling enables and promotes the thorough documentation proposed in prior work (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019; Pushkarna et al., 2022)

3 Empirical Analysis of Data Provenance

3.1 Licenses in the Wild

This work constitutes the first extensive study of empirical license use for Natural Language Processing datasets. In this section, we share the insights we have gathered from our large-scale annotation and categorization. There is an important assumption in this section: the OpenAI Terms of Use is a contract, not a license, which prohibits the development of competing models using its outputs. For simplicity, we treat this as a Non-Commercial license in our analysis, though this is disputed for third parties who did not generate the OpenAI data themselves and therefore may not be bound by their terms (see Section 4 for discussion). Given the intention of OpenAI not to facilitate competitive commercial uses, we follow their categorization for this analysis.

Frequency of license types Figure 2 shows the distribution of licenses. The most common licenses are CC-BY-SA 4.0 (15.7%), the OpenAI Terms of Use (12.3%), and CC-BY 4.0 (11.6%). While most licenses are common and recognizable, there is a long tail of variants with unique settings, as well as a large set of Custom licenses accounting for 9.6% of all recorded licenses on their own. **This wide license diversity illustrates the challenge to startups and less resourced organizations attempting to navigate responsible training data collection, its legality and ethics.**
















CORRECT LICENSE		LICENSE ACCORDING TO AGGREGATORS (AGG.)				
LICENSE	COUNT	AGG.	COMM.	UNSPEC.	NON-COMM.	ACAD.-ONLY
Commercial	856 (46.1%)		349	507	0	0
			176	677	1	2
			313	520	1	22
Unspecified	570 (30.7%)		112	458	0	0
			164	395	6	5
			31	523	1	15
Non-Commercial	352 (19.0%)		49	303	0	0
			113	152	80	7
			2	191	157	2
Academic-Only	80 (4.3%)		9	71	0	0
			9	65	2	4
			5	65	2	8
Total	1858 (100%)		519 (28%)	1339 (72%)	0 (0%)	0 (0%)
			462 (25%)	1289 (69%)	89 (5%)	18 (1%)
			351 (19%)	1299 (70%)	161 (9%)	47 (3%)

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

Distribution of Restrictive Licenses In total, 85% of dataset licenses request attribution, and 30% include a share alike clause.[‡] Datasets which request attribution pose challenges for practitioners who commonly train on hundreds of datasets and either don’t cite them at all (OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023) or simply cite an aggregation of data, which often falls short of the license’s conditions of attributing the specific repository or paper. Furthermore, “Share alike” clauses poses challenges for practitioners re-packaging data collections usually with multiple conflicting share-alike licenses without a clear way to resolve them (like Longpre et al. (2023a); Wang et al. (2022c) and others in the DPCollection). Frequently, practitioners will over-write share-alike licenses with more restrictive or even less restrictive conditions.

Missing or Unspecified Licenses. Next, we compare our manually reviewed licensing terms, to the licenses for the same datasets, as documented in the aggregators GitHub, HuggingFace, and Papers with Code. Table 2 shows that these crowdsourced aggregators have an extremely high proportion of missing (“Unspecified”) licenses, ranging from 69-72%, as compared to our protocol which yields only 30% “Unspecified”. The problem with “Unspecified” licenses is that it is unclear whether it is due to a shortcoming of the aggregator or because creators intentionally released them without a license. Consequently, risk-averse developers are forced to avoid many valuable datasets, which they would use otherwise if they were given assurance that there is indeed no license. As part of DPCollection, we manually reassign 46-65% of dataset licenses (depending on the platform), resulting in much higher coverage, thus giving risk-averse developers more confidence and breadth in their dataset utilization.

Incorrectly Specified Licenses. Table 2 also finds real licenses as assigned by us are frequently stricter than the ones by aggregators. GitHub, Hugging Face and Papers with Code each label license 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 licenses attached to code in GitHub repositories for licenses attached to data.

[‡]“Share alike” is a copyright term meaning adaptations or copies of a work are required to be released under the same license as the original.

METRICS	COMMERCIAL		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	0.52	1.2±0.0	0.16	1.1±0.0	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 TEXT LENGTHS	1043.4±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	1580.7±965.6	5.37
SYNTHETIC	12.8%±2.1	-	13.6%±1.7	-	45.5%±3.4	-

Table 3: The mean number of features (e.g. tasks or languages) per dataset, and the mean entropy of the distribution, representing the diversity of categories. **Non-Commercial / 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.

3.2 How does Data Availability Differ by License Use Category?

While non-commercial and academic-only licenses 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 license restrictions and see how they differ. Specifically, we ask: *Does complying with licenses 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 licenses: (1) Commercially viable, (2) Non-Commercial/Academic-Only (NC/A-O), or (3) Unspecified license. We group together Non-Commercial and Academic-Only conditions as the distinction will rarely matter for developers. We argue in Section 4 that datasets without any license (Unspecified) have not imposed any conditions, so can often be treated as commercially viable, but this may depend on a developer’s risk tolerance and jurisdiction.

Non-Commercial & Academic-Only Licensed Datasets have statistically greater diversity in their representation of tasks, topics, sources, and target text lengths. For each of these features, Table 3 illustrates the mean number per dataset, broken down by license 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. Figure 4 shows where this diversity comes from. The most NC/A-O task categories include Brainstorming, Explanation, Logic & Math, 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 significantly higher for NC/A-O datasets than commercial 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 license categories are similar on average, the target text lengths are significantly higher for NC/A-O datasets (103 vs 677). This breakdown is further illustrated in Figure 5, 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, as compared to < 14% in more permissive license categories. Taori et al. (2023); Wang et al. (2022a); Xu et al. (2023a) 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

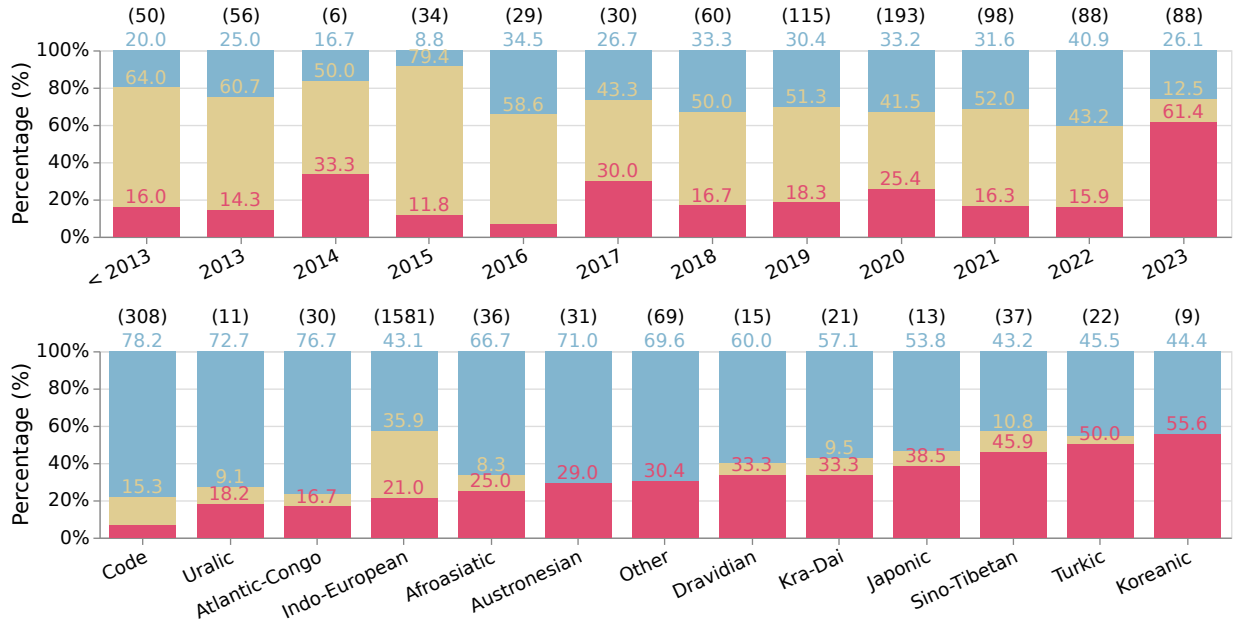


Figure 3: The distribution of datasets in each time of collection (top) and language family (bottom) category, with total count above the bars, and the portion in each license use category shown via bar color. **Red** is Non-commercial/Academic-Only, **Yellow** is Unspecified, and **Blue** is Commercial.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

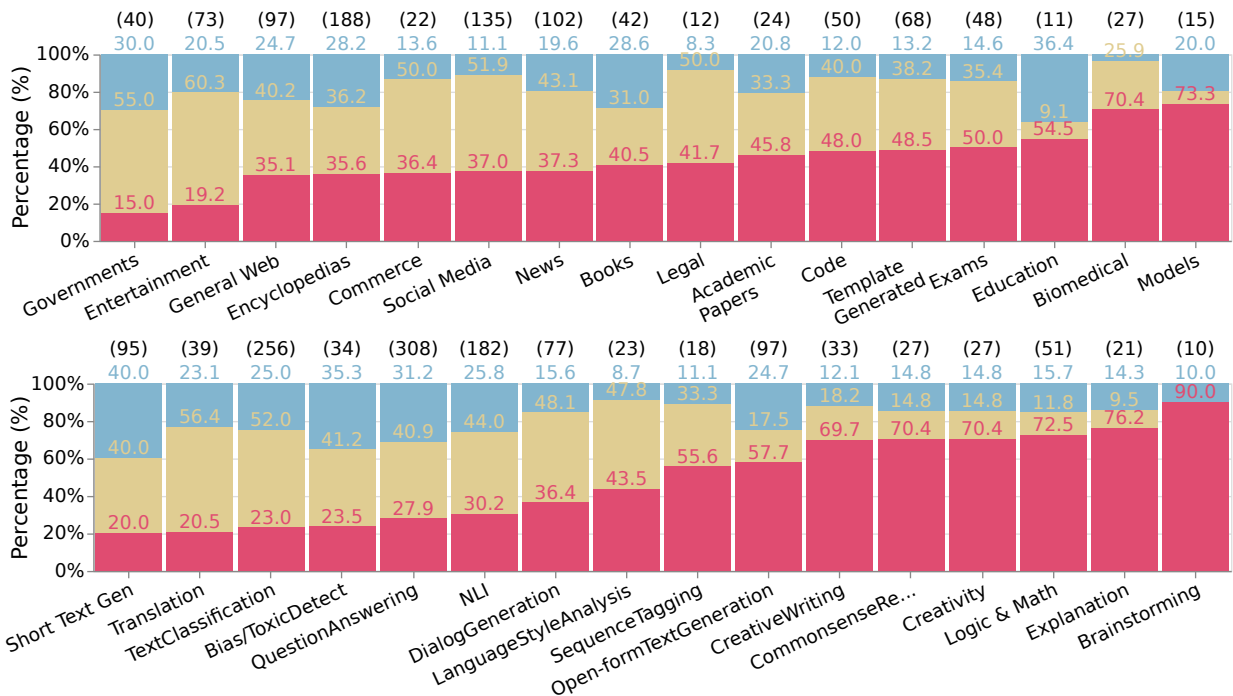


Figure 4: The distribution of datasets in each **Domain Source (top)** and **task (bottom)** category, with total count above the bars, and the portion in each license use category shown via bar color. **Red** is Non-commercial/Academic-Only, **Yellow** is Unspecified, and **Blue** is Commercial. **Creative,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

evident from the concentration of creative, brainstorming, and reasoning tasks baked into them, as 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. The OpenAI Terms of Use state “you may not...use output from the Services to develop models that compete with OpenAI.” which we discuss in Section 4.[§]

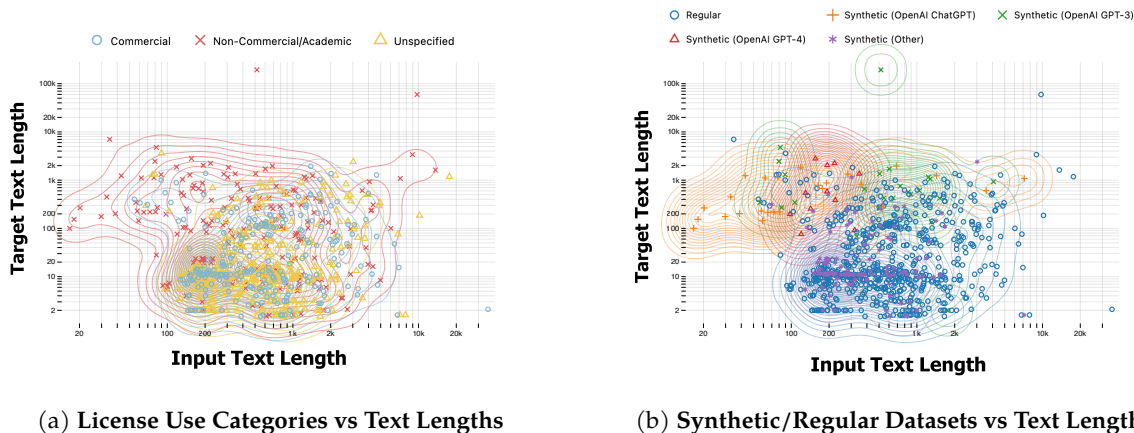


Figure 5: Across finetuning datasets, we visualize their mean input (x-axis) and target (y-axis) text lengths, measured in log-scaled number of words. The colors indicate either their license use category (left) or whether they were machine generated or human collected (right). **Long target texts are represented in large part by Non-Commercial and Synthetic datasets, that are often generated by commercial APIs.**

2023 has a large spike in license usage, and in NC/A-O licensed data, representing 61%, as compared to 20% on average in prior years. 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 Figure 3, prior to 2023, no year saw greater than 1/3 of the datasets released as NC/A-O. However, in 2023, which includes many of the most popular and diverse datasets, the NC/A-O rate is 61%. Furthermore, most datasets were unaccompanied by a license prior to 2022 (50-80%), as compared to only 12% in 2023. The shift to more license use, and more restrictively conditioned data releases may foretell future challenges to open data, if the trend continues.

Commercial datasets have greater language variety, but low-resource language datasets see the least commercial coverage. Table 3 shows that commercial datasets actually have greater diversity of languages than NC/A-O. However, when broken down by language family, as in Figure 3, 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 licensed repositories. Interestingly, 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.

3.3 Broader Characteristics of the Data

In addition to understanding systematic differences in the data by license, 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.comingsoon.com, for further research and examination.

[§]<https://openai.com/policies/terms-of-use>

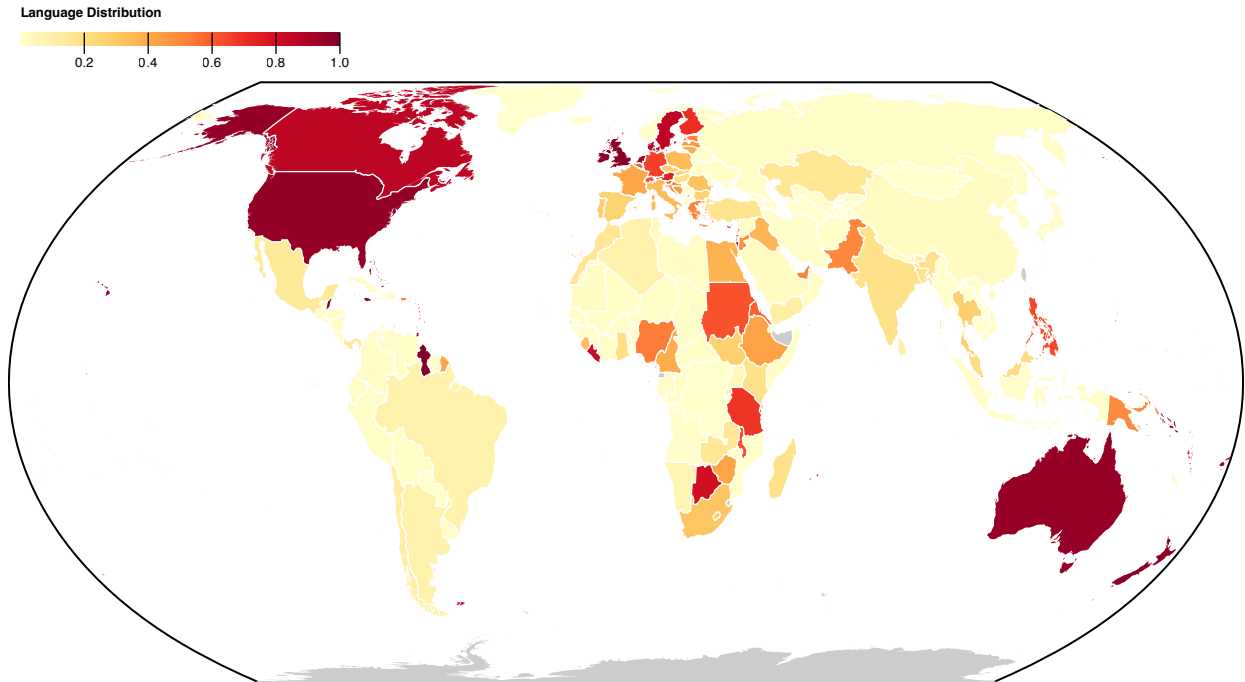


Figure 6: A global heatmap measuring how well each country’s spoken languages are represented by the composition of natural language datasets in DPCollection, as calculated by Section 3.3. **English-speaking and Western European nations are best represented, while the Global South sees limited coverage.**

Language representation is heavily skewed to English and Western European Languages. Following Talat et al. (2022)’s recommendations in data transparency and documentation in demographic analysis, and corroborating Kreutzer et al. (2022)’s similar analysis for pretraining corpora, we find a stark Western-centric skew in representation. Figure 6 illustrates the coverage per country according to the spoken languages and their representation in DPCollection. 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} which is a binary indicator that is 1 if dataset $i \in D$ contains language l and 0 otherwise.

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

The distribution visualized in Figure 6 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, according to Section 3.3, the text source and dialect of the language contained in these datasets almost always originates from North American or European creators and web sources (though this is difficult to measure precisely). These observations corroborate similar findings in the geo-diversity of image data in the vision domain (Shankar et al., 2017; De Vries et al., 2019; Mahadev and Chakravarti, 2021). The resulting models trained on these datasets are likely to have inherent bias, underperforming in critical ways for users of models outside of the west (Ahia et al., 2021).

The primary drivers of dataset curation are Academic organizations, supplying 69%, followed by 21% industry labs, and 17% research institutions. These metrics describe the scale of dataset curation contributions, but not the influence each dataset has had on the community. Table 4a 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.

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%
NATURAL LANGUAGE INF...	21.1%
Textual Entailment	14.6%
Natural Language Infer...	5.3%
Fact Verification	1.3%
OPEN-FORM TEXT GENER...	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-Choi...	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 MATHEMAT...	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

Table 4: A summary of the distribution of **Creators**, **Topics**, and **Source Domains** across all 1800+ datasets. Datasts can have multiple creators, text topics, and sources.

Text datasets focus on topics of Language & Linguistics, General Knowledge, Logic, & Lifestyle. Prior data collection work focuses predominantly on describing datasets by their task compositions (Sanh et al., 2021; Wang et al., 2022a; Longpre et al., 2023a), but rarely by their actual topics (except (Gao et al., 2020) in their Appendix). Table 4b shows the most popular topics, clustered by category, with their representation across datasets. Like most NLP tasks, much of this text data focuses on communication and language understanding topics, followed closely by general knowledge, routine, sports, and education.

Text datasets are sourced primarily from Online Encyclopedias (22%), Social Media (16%), scraped from the General Web (11%), News (11%), Entertainment web resources (9%). While practitioners document their individual dataset sources in their published papers, this information is unstructured and can be hard to find. As a result, massive collections of widely used datasets rarely compile the distribution of their original sources, instead just citing the papers. After a series of dataset compilations and re-packaging, the original sources are often lost or not well known. By manually scanning approximately 500 academic papers our volunteers annotated the original text sources and compiled them into domain clusters, to permit attribution and analysis, as summarized in Table 4c. Among the individual most adopted sources by the used sources are wikipedia.org (14.9%), undisclosed webpage scrapes (7.0%), reddit (6.2%), and Twitter (4.0%). The least represented domains are Commerce, Reviews, Legal, Academic Papers, and Search Queries, among others.

4 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 licenses are in tension, major legal ambiguities remain in data licensing.

Background Copyright laws aim to encourage written and artistic expression by giving authors exclusive rights to copy, distribute, and adapt their work (Patterson, 2003; Burger, 1988). Open-source licenses first emerged as legal tools to encourage collaboration around software development (Von Krogh and Von Hippel, 2003). A range of licenses with different terms and purposes exists including the MIT License, Creative Commons Licenses, and the Apache License, as well as the newer Responsible AI License (RAIL) and AI2 ImpACT Licenses.[¶] The interplay between copyright and licenses can be understood in the following way: copyright automatically gives creators exclusive rights in their works and creators assign these rights to others through license agreements. As we will explore, the open-source licenses that emerged in the last three decades are not always well-equipped to handle the unique characteristics of data, and especially supervised AI training data. Meanwhile, it remains unclear how relevant laws, including those related to copyright and fair use, should be applied to the unique challenges raised by Generative AI and supervised datasets (Lee et al., 2023). In this section, we highlight some of the key legal challenges and ambiguities related to supervised datasets.

Lifecycle of a dataset We focus on *supervised datasets*, which we define as datasets that are created for machine learning (mainly for finetuning and alignment) and where dataset creators made copyrightable contributions in the form of annotations or compilations. A typical supervised dataset is the result of a process that involves several stages of scraping (or machine generation) and annotation by different entities. Generally, raw data is created by people interacting with internet platforms, such as individuals writing articles, sharing artworks, or engaging in online discussion forums. The copyrights to this raw data are normally held by individual users (e.g. Reddit) or by the platform (e.g. Amazon Reviews). Much of this data has been scraped to construct unsupervised datasets for machine learning and this use is commonly justified on the basis of fair use or data mining exceptions to copyright (Henderson et al., 2023; Sobel, 2017; Lee et al., 2023; Samuelson, 2023; Lemley and Casey, 2020). However, we find that many common supervised

[¶]See <https://www.licenses.ai/blog/2023/3/3/ai-pubs-rail-licenses> and <https://allenai.org/impact-license#licenses>. These license templates propose terms aimed at encouraging more responsible or risk-based machine learning practices, see also Contractor et al. (2022)

datasets are generated by annotating small samples of scraped raw data using human annotators or large language models. The annotated data is then published with a license agreement. In stark contrast to the copyrighted content that is scraped from the web, supervised datasets were created for the sole purpose of furthering machine learning. The focus of the legal discussion in this section is on how supervised dataset creators can constrain the usage of the copyrightable content they create through licenses and other legal mechanisms. Though we do not address them here, there are several important related questions on the use of copyrighted works to create supervised datasets and on the copyrightability of training datasets.

Supervised Dataset Example: SQuAD

Rajpurkar et al. (2016) present a prototypical supervised dataset on reading comprehension. To create the dataset, the authors take paragraph-long excerpts from 539 popular Wikipedia articles and hire crowd-source workers to generate over 100,000 questions whose answers are contained in the excerpt. For example:

Wikipedia Excerpt *In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.*

Worker-generated question: What causes precipitation to fall? **Answer:** Gravity

Here the authors use Wikipedia text as a basis for their data and their dataset contains 100,000 new question-answer pairs based on these texts.

Copyright laws vary by jurisdiction and are subjective, so it is challenging to develop technical safeguards that guarantee compliance. The legal analysis surrounding supervised datasets is complicated by the lack of a uniform global legal framework to address copyright concerns. Different jurisdictions have different and evolving laws. Therefore, the location of model developers and training data creators as well as where and when data was collected may influence the legal analysis. For example, the United States has a fair-use exception to copyright that allows the limited use of copyrighted material under certain circumstances without requiring permission from the rights holders (17 U.S.C. §107). The EU has no fair-use provision but does have an explicit copyright exception to allow data mining under certain conditions, like obtaining lawful access to the data (Margoni and Kretschmer, 2022). Meanwhile, datasets themselves generally enjoy copyright protection in the U.S. (Lee et al., 2023) while the E.U. recently created a unique set of rights for dataset creators with the purpose of incentivizing research and development related to databases (Derclaye and Husovec, 2022). In addition to differences across jurisdictions, there are also several international agreements related to copyright Ricketson and Ginsburg (2022). Ultimately, it can be challenging to determine which laws should apply to a given machine learning project when the relevant rules vary between the locations where the data was scraped and annotated, where it was downloaded, where the model was trained, and where the model was deployed.

While geographical disparities in regulatory frameworks present one set of challenges, the subjectivity inherent in determining whether copyright infringement has occurred makes it even more challenging to design technical safeguards. For example, in the U.S. part of the copyright infringement analysis depends on whether two works are subjectively similar from the perspective of an ordinary person (Mohler, 1999; Cohen, 1986; Balganesch et al., 2014). This is a subjective standard and existing case law may be challenging to extend to generative AI outputs. As a result, while there are technical strategies that can reduce the risk of infringement (Henderson et al., 2023; Sag, 2023; Vyas et al., 2023), it will be difficult for developers to create technical safeguards that eliminate this risk entirely.

Open legal question regarding copyright and model training. Apart from these jurisdictional and interpretive ambiguities, the process of training a model raises specific copyright questions (Epstein et al., 2023). 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 scraping 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. If the scraped data is protected by copyright, then creating training data corpora may raise copyright issues (Quang, 2021). Second, copyright holders generally have an exclusive right to create derivative works (e.g., translations of a work) but it is not clear whether a trained machine learning model should be considered a derivative of the training data (Lee et al., 2023). If models are considered to be derivative works, then training a model would be more likely to violate the rights of the training data’s copyright holders (Gervais, 2021).

In the U.S., the fair use exception may allow models to be trained on protected works (Henderson et al., 2023; Lemley and Casey, 2020; Sobel, 2017; Samuelson, 2023). As these authors explain, the training of machine learning models on copyrighted content may be permissible if the underlying works are significantly “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 prior works. As the authors above highlight, the application of fair use in the context is still evolving and several of these issues are currently being litigated (see e.g., *Andersen v. Stability*, *Doe v. GitHub*, and *Tremblay v. OpenAI*).

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. The prior literature on fair use and machine learning tends to focus on copyrighted art or text that was scraped to train a model. These scraped 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 Henderson et al. (2023) and Sobel (2017), the fair use analysis depends in part on whether a trained model copies the “expressive purpose” of the original work. 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 license agreement govern the subsequent use of this data. However, it is rare in practice for an 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 license terms of many popular dataset collections are conflicting.

Licenses used for datasets are often ill-suited for this purpose. Beyond the intricate interplay between training data and fair use, the frequently misapplied licensing frameworks for datasets present another set of complications. Most open-source licenses were designed for software, but we find them being attached to datasets. These licenses were intended to be applied to software, not data, which creates challenges (Meeker, 2022). One of the challenges is that licenses like the Apache and the Creative Commons outline restrictions related to “derivative” or “adapted works” but it remains unclear if a trained model should be classified as a derivative work. This issue is further exacerbated when multiple datasets, each potentially governed by a different open-source license, are amalgamated into collections. If the requirements of the underlying license agreements are irreconcilable, such as different copyleft requirements, this makes it extremely hard for developers to use certain collections while respecting all license terms. To remedy these issues, new licenses are being proposed to address the needs of machine learning datasets such as the BigScience Responsible AI License or an adaptation of the MIT License that requires additional permissions for model training proposed by Ioannidis et al. (2023). Despite these new proposals, we find that the majority of datasets are licensed under conventional open-source licenses.

[¶]Data ownership and data copyright are complex topics (Ginsburg, 1992). We assume that the creators of supervised datasets have some form of copyright in their dataset, though there is often content in these datasets that is owned by third parties. If they satisfy the requirements for copyrightability, dataset creators would have a copyright interest in any new content they create (e.g. annotations). In the U.S., datasets themselves may also be copyrightable as compilations (Lee et al., 2023) while the E.U. provides so-called *sui generis* rights for databases (Derclaye and Husovec, 2022).

LLM-generated annotations raise additional legal considerations 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”^{**}. 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. On the one hand, publishing such a dataset does not directly compete with OpenAI. On the other hand, it seems foreseeable that such a dataset could enable third parties (who did not themselves use OpenAI) to create competing LLMs. In the U.S., there are several doctrines of secondary or indirect copyright liability aimed to enforce copyright in cases where there is no direct infringement (Grossman, 2005; Lee et al., 2023). 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 (Marks and Moll, 2023). Time will tell the extent to which OpenAI and other LLM service providers can enforce their terms of use against third parties. However, a prominent researcher at Google has already resigned citing concerns that OpenAI outputs were used to train BARD (Victor and Efrati, 2023) In light of these legal ambiguities, our tool gives developers the ability to exclude OpenAI-generated datasets.

While legal issues remain ambiguous, practitioners are making decisions on data use and model training. In the face of these pervasive legal uncertainties, practitioners’ decisions regarding data usage are ultimately guided by a blend of factors including the specific licensing terms, the origin of datasets, and the degree of usage of a given dataset by others. Navigating this landscape requires striking a delicate balance between risk mitigation and the need for sufficient resources. This equation, however, varies across regions, applications, and corporate environments, influenced by factors such as competition, risk, and regional legislation. A strategy for partially mitigating these uncertainties is for model providers to indemnify users, as done by Google Cloud Suggs and Venables (2023). However, this may not be feasible for resource-constrained developers and, while it protects end-users, it does not solve the issues faced by model developers or dataset curators.

Our Approach. The fundamental purpose of copyright is to encourage creativity and innovation. As we highlighted in the sections above, the current legal landscape remains ambiguous and this lack of clarity can stifle innovation as developers fear legal repercussions. Through our audit and tooling, we seek to provide important information for practitioners to make informed decisions in an otherwise ambiguous landscape, guided by their own legal interpretation and risk tolerance. This information includes data license lineages, a categorization of license terms, details on data creators, and the underlying data sources (e.g. web or LLM). In light of ongoing litigation and a lack of legal certainty, we attempted to give developers In creating a repository of data licensing information, we are also taking a step towards encouraging dataset creators to be more thoughtful about the licenses that they select. Dataset creators are well-positioned to understand the appropriate uses of the datasets they publish and licenses can be a tool to communicate these restrictions and to encourage responsible AI development. We further aim to highlight that machine learning practitioners should take dataset license terms seriously, as they may have real impacts on how their models may be used in practice. Ultimately, thoughtful data licensing could be leveraged to promote more responsible, inclusive, and transparent machine learning practices.

NOTICE: Collected License Information is NOT Legal Advice. It is important to note we collect *self-reported* licenses, and categorize them according to our best efforts, as a volunteer research and transparency initiative. The information provided by any of our works and any outputs of the Data Provenance Initiative do not, and are not intended to, constitute legal advice; instead, all information, content, and materials are for general informational purposes only. Readers and users should seek their own legal advice from counsel in their relevant jurisdiction.

^{**}<https://openai.com/policies/terms-of-use>

5 Related Work

Data Documentation A long line of work has highlighted the importance of data and its documentation in natural language processing (Paullada et al., 2021; Rogers, 2021; Meyer et al., 2023; Gururangan et al., 2018; Muennighoff et al., 2023b). In particular, these works stress the challenges posed by poor documentation to reproducibility, good science, and generally well-understood model behavior (Sambasivan et al., 2021a; Bandy and Vincent, 2021; Longpre et al., 2023b). Recent work has also explored the importance of documenting AI ecosystems (Bommasani et al., 2023b) and the supply chain from data to models (Cen et al., 2023).

Data Analysis and Exploration Several notable works have conducted large-scale analyses into data, particularly pretraining text corpora (Gao et al., 2020; Dodge et al., 2021; Kreutzer et al., 2022; Laurençon et al., 2022; Scao et al., 2022a,b; McMillan-Major et al., 2022). Other works have investigated the geo-diversity of vision-based datasets (Shankar et al., 2017; De Vries et al., 2019; Mahadev and Chakravarti, 2021). Different forms of data governance have been proposed to centralize responsibility and documentation over datasets, including for the BigScience project (Jernite et al., 2022) and a Public Data Trust (Chan et al., 2023). In terms of finding and visualizing datasets, a few recent tools have been proposed (Färber and Leisinger, 2021; Viswanathan et al., 2023).

Transparency and accountability Adjacent to the realm of legality, prior works have strongly advocated and provided frameworks for documentation and audits to increase transparency and accountability in AI systems (Miceli et al., 2022; Kapoor et al., 2023; Raji and Buolamwini, 2022). In a manner akin to DPI, which draws upon the collective knowledge of legal and machine learning experts, earlier research has also underscored the significance of interdisciplinary collaborations (Hutchinson et al., 2021). Datasheets for datasets Gebru et al. (2021) and Data Statements Bender and Friedman (2018) both provide structured frameworks for revealing essential metadata such as the motivation behind intended use. Pushkarna et al. (2022) expanded on datasheets with “Data Cards” for sources, collection, ethics, and adoption.

Similarly, Mitchell et al. (2019) introduced model cards to benchmark model performance across demographic groups and disclose evaluation procedures. Crisan et al. (2022) proposed interactive model card as an alternative mode of documentation and metadata sharing. Complementary to transparency regarding the dataset’s creation process, Corry et al. (2021) provide a framework that guides users on how to navigate datasets as they approach the end of their life-cycle. DPI builds upon the foundational frameworks laid out in these earlier studies, with a specific focus on addressing the licensing aspects of dataset curation. Our goal is to equip users with a comprehensive understanding of the legal risks associated with dataset usage.

Dataset legality The legality of the datasets used to train large base models has recently received significant attention (Sag, 2020; Henderson et al., 2023). The challenge of determining the legality of employing different datasets becomes particularly complex due to the intricate nature of dataset creation processes. Lee et al. (2023) break up the stages of dataset creation and model generation and assess the relevant copyright questions in the US legal system. These processes often involve multiple licenses and restrictions that can interact in ways that obscure the final legal risk. Soh (2021) propose a high-level framework for pinpointing the areas within dataset creation and usage where legal analysis is necessary, but do not apply this framework to any existing datasets. Min et al. (2023) demonstrate that refraining from training on copyrighted or highly restricted datasets has a detrimental impact on downstream performance. Their proposed solution involves using a language model trained on “low-risk” text and augmenting it with a data-store containing “high-risk” text which can be modified appropriately as the legal landscape clarifies over time. (Lee et al., 2023) DPI enhances these investigations by involving legal experts in the development of a framework for assessing a dataset’s “risk” and annotating the “risk” associated with numerous existing high-profile datasets.

Acknowledgements

We would like to thank Katherine Lee, A. Feder Cooper, Peter Henderson, Aviya Skowron and Stella Biderman for valuable comments and feedback.

References

- Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.282. URL <https://aclanthology.org/2021.findings-emnlp.282>.
- Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. <https://github.com/nomic-ai/gpt4all>, 2023.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- Anthropic. Model card and evaluations for claude models. 2023.
- Arstechnica. Stable diffusion copyright lawsuits could be a legal earthquake for ai, 2023. URL <https://arstechnica.com/tech-policy/2023/04/stable-diffusion-copyright-lawsuits-could-be-a-legal-earthquake-for-ai/>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Shyamkrishna Balganesh, Irina D Manta, and Tess Wilkinson-Ryan. Judging similarity. *Iowa Law Review*, 100: 267, 2014.
- Jack Bandy and Nicholas Vincent. Addressing “documentation debt” in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.
- Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.
- Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6: 587–604, 2018. doi: 10.1162/tacl_a_00041. URL <https://aclanthology.org/Q18-1041>.
- Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*, 2022.
- Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index, 2023a.
- Rishi Bommasani, Dilara Soyulu, Thomas Liao, Kathleen A. Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models. *ArXiv*, abs/2303.15772, 2023b. URL <https://arxiv.org/abs/2303.15772>.
- Rishi Bommasani, Dilara Soyulu, Thomas I Liao, Kathleen A Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models. *arXiv preprint arXiv:2303.15772*, 2023c.

- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- Peter Burger. The berne convention: Its history and its key role in the future. *Journal of Law and Technology*, 3: 1, 1988.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- Sarah H. Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray. Ai supply chains (and why they matter), April 2023. URL <https://aipolicy.substack.com/p/supply-chains-2>. The second post in our series On AI Deployment.
- Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. Reclaiming the digital commons: A public data trust for training data. *arXiv preprint arXiv:2303.09001*, 2023.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- Amy B Cohen. Masking copyright decisionmaking: The meaninglessness of substantial similarity. *UC Davis Law Review*, 20:719, 1986.
- Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm. <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>, 2023.
- Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 778–788, 2022.
- Frances Corry, Hamsini Sridharan, Alexandra Sasha Luccioni, Mike Ananny, Jason Schultz, and Kate Crawford. The problem of zombie datasets: A framework for deprecating datasets. *ArXiv*, abs/2111.04424, 2021. URL <https://arxiv.org/abs/2111.04424>.
- Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, 2022.
- Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019.
- Estelle Derclaye and Martin Husovec. Sui generis database protection 2.0: judicial and legislative reforms. *European Intellectual Property Review*, 44(6):323–331, 2022.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.
- Jon Durbin. Airoboros: Using large language models to fine-tune large language models. <https://github.com/jondurbin/airoboros>, 2023.

- Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.113. URL <https://aclanthology.org/2021.eacl-main.113>.
- Ronen Eldan and Yuanzhi Li. Tinstories: How small can language models be and still speak coherent english?, 2023.
- Ziv Epstein, Aaron Hertzmann, Laura Herman, Robert Mahari, Morgan R Frank, Matthew Groh, Hope Schroeder, Amy Smith, Memo Akten, Jessica Fjeld, et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.
- Kawin Ethayarajh, Heidi Zhang, Yizhong Wang, and Dan Jurafsky. Stanford human preferences dataset, 2023. URL <https://huggingface.co/datasets/stanfordnlp/SHP>.
- Michael Färber and Ann-Kathrin Leisinger. Datahunter: A system for finding datasets based on scientific problem descriptions. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 749–752, 2021.
- Jonathan Frankle. Tweet by mosaic ml. <https://twitter.com/jefrankle/status/1654848529834078208>, 2023.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. 2022.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. Blog post, April 2023. URL <https://bair.berkeley.edu/blog/2023/04/03/koala/>.
- Daniel J Gervais. Ai derivatives: The application to the derivative work right to literary and artistic productions of ai machines. *Seton Hall L. Rev.*, 52:1111, 2021.
- Jane C Ginsburg. No sweat copyright and other protection of works of information after feist v. rural telephone. *Columbia Law Review*, 92:338, 1992.
- Craig A Grossman. 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 Law Review*, 53:141, 2005.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL <https://aclanthology.org/N18-2017>.
- Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor, 2022.
- Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445918. URL <https://doi.org/10.1145/3442188.3445918>.
- Dimitrios Ioannidis, Jeremy Kepner, Andrew Bowne, and Harriet S Bryant. Are chatgpt and other similar systems the modern lernaean hydras of ai? *arXiv preprint arXiv:2306.09267*, 2023.
- Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, et al. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2222, 2022.
- Sayash Kapoor, Emily F. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica R. Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russel A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M Stewart, Gilles Vandewiele, and Arvind Narayanan. Reforms: Reporting standards for machine learning based science. *ArXiv*, abs/2308.07832, 2023. URL <https://arxiv.org/abs/2308.07832>.
- Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning, 2023.
- Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buczaczyński, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. *ArXiv*, abs/2301.10140, 2023.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations—democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.
- Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization, 2022.
- Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. Longform: Optimizing instruction tuning for long text generation with corpus extraction, 2023.
- Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Froberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna

- Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.
- Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin’bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023.
- Mark A Lemley and Bryan Casey. Fair learning. *Texas Law Review*, 99:743, 2020.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, 2021.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society, 2023a.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umaphathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023b.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023a.
- Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023b.
- Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.
- Rohan Mahadev and Anindya Chakravarti. Understanding gender and racial disparities in image recognition models. *arXiv preprint arXiv:2107.09211*, 2021.
- Thomas Margoni and Martin Kretschmer. A deeper look into the eu text and data mining exceptions: harmonisation, data ownership, and the future of technology. *GRUR International*, 71(8):685–701, 2022.
- Colin P. Marks and Douglas K. Moll. *The Law of Business Torts and Unfair Competition: Cases, Materials, and Problems*. American Casebook Series. West Academic, 2023. ISBN 9781647084905. URL <https://books.google.com/books?id=K1fXzwEACAAJ>.

- Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, et al. Documenting geographically and contextually diverse data sources: The bigscience catalogue of language data and resources. *arXiv preprint arXiv:2201.10066*, 2022.
- Heather Meeker. Beyond open data: The only good license is no license. *PLI Chronicle: Insights and Perspectives for the Legal Community*, April 2022.
- Anna P. Meyer, Aws Albarghouthi, and Loris D’Antoni. The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23*, page 193–204, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593988. URL <https://doi.org/10.1145/3593013.3593988>.
- Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. Documenting data production processes: A participatory approach for data work. volume 6, New York, NY, USA, nov 2022. Association for Computing Machinery. doi: 10.1145/3555623. URL <https://doi.org/10.1145/3555623>.
- Sewon Min, Suchin Gururangan, Eric Wallace, Hanna Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. Silo language models: Isolating legal risk in a nonparametric datastore. *ArXiv*, abs/2308.04430, 2023. URL <https://arxiv.org/abs/2308.04430>.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.
- Jarrold M Mohler. Toward a better understanding of substantial similarity in copyright infringement cases. *U. Cin. L. Rev.*, 68:971, 1999.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.
- Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*, 2023a.
- Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023b.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- Huu Nguyen, Sameer Suri, Ken Tsui, and Christoph Schuhmann. The open instruction generalist (oig) dataset. <https://laion.ai/blog/oig-dataset/>, 2023.
- OpenAI. Gpt-4 technical report, 2023.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL <https://arxiv.org/abs/2203.02155>.
- Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis, 2023.

- L Patterson. Copyright in 1791: An essay concerning the founers' view of the copyright power granted to congress in article i, section 8, clause 8 of the us constitution. *Emory Law Journal*, 52:909, 2003.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. URL <https://arxiv.org/abs/2306.01116>.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. *arXiv preprint arXiv:2305.15425*, 2023.
- Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. On the challenges of using black-box apis for toxicity evaluation in research, 2023.
- Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023.
- Jenny Quang. Does training ai violate copyright law? *Berkeley Tech. LJ*, 36:1407, 2021.
- Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial ai products. *Communications of the ACM*, 66(1): 101–108, 2022.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- Sam Ricketson and Jane C. Ginsburg. *International Copyright and Neighboring Rights: The Berne Convention and Beyond*. Oxford University Press, August 2022.
- Anna Rogers. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.170. URL <https://aclanthology.org/2021.acl-long.170>.
- Matthew Sag. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023.
- Matthew J. Sag. The new legal landscape for text mining and machine learning. In *Journal of the Copyright Society of the USA*, 2020.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL <https://doi.org/10.1145/3411764.3445518>.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. In *CHI, CHI '21*, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL <https://doi.org/10.1145/3411764.3445518>.

Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *ICLR 2022*, 2021. URL <https://arxiv.org/abs/2110.08207>.

Joseph R. Saveri, Cadio Zirpoli, Christopher K.L. Young, and Kathleen J. McMahon. Paul tremblay, mona awad vs. openai, inc., et al., 2023. URL https://storage.courtlistener.com/recap/gov.uscourts.cand.414822/gov.uscourts.cand.414822.1.0_1.pdf. Case 3:23-cv-03223-AMO Document 1 Filed 06/28/23, UNITED STATES DISTRICT COURT, NORTHERN DISTRICT OF CALIFORNIA, SAN FRANCISCO DIVISION.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022a.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*, 2022b.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Damien Sileo. tasksource: A dataset harmonization framework for streamlined nlp multi-task learning and evaluation. *arXiv*, abs/2301.05948, 2023.

Benjamin LW Sobel. Artificial intelligence’s fair use crisis. *Columbia Journal of Law & the Arts*, 41:45, 2017.

Jerrold Soh. Building legal datasets. *ArXiv*, abs/2111.02034, 2021. URL <https://arxiv.org/abs/2111.02034>.

Spacerini. Gaia search tool. 2021. URL <https://huggingface.co/spaces/spacerini/gaia>.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *NeurIPS*, 2020.

Neal Suggs and Phil Venables. Protecting customers with generative AI indemnification, 2023. URL <https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification>.

Zeeraq Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, 2022.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutvi Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vercel. Sharegpt, 2023. URL <https://sharegpt.com/>.

- Jon Victor and Amir Efrati. Alphabet’s google and deepmind pause grudges, join forces to chase openai. *The Information*, 2023.
- Vijay Viswanathan, Luyu Gao, Tongshuang Wu, Pengfei Liu, and Graham Neubig. Datafinder: Scientific dataset recommendation from natural language descriptions. *arXiv preprint arXiv:2305.16636*, 2023.
- Georg Von Krogh and Eric Von Hippel. Special issue on open source software development, 2003.
- Nikhil Vyas, Sham Kakade, and Boaz Barak. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022a. URL <https://arxiv.org/abs/2212.10560>.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022b. URL <https://arxiv.org/abs/2204.07705>.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, 2022c.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.
- Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, 2021.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015.
- Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, 2021.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions, 2023a.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data, 2023b.
- Joanna Yoo, Kuba Perlin, Siddhartha Rao Kamalakara, and João G. M. Araújo. Scalable training of language models using jax pjit and tpuv4, 2022.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.

Appendix

A Contributors

Here we enumerate the author contributions. We would like to 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.

- **Shayne Longpre** Core Contributor. Primary designer and coder of the repository and explorer interface. Led audit implementation, and analysis, as well as the manual annotation process.
- **Robert Mahari** Core Contributor. Led the legal analysis, and licensing annotation design.
- **Anthony Chen** Core Contributor. Led automatic inferencing of dataset text metrics, topics, and task category annotations. Supported writing, analysis, and code testing.
- **Naana Obeng-Marnu** Core contributor. Led visualization design, particularly interactive visualizations in the Data Provenance Explorer.
- **Damien Sileo** Core contributor. Led data aggregator linking, and metadata scraping. Supported writing, analysis, source annotation and adding datasets.
- **William Brannon** Core contributor. Added 8 data collections, supported writing and data analysis.
- **Niklas Muennighoff** Core contributor. Added several large data collections, supported writing, analysis, visualization, and source annotations.
- **Nathan Khazam** Core contributor. Led licensing annotation effort and supported adding datasets along with testing.
- **Jad Kabbara** Core contributor and advisor. Led text source annotation effort and supported with framing, writing and analysis.
- **Kartik Perisetla** Core contributor. Added several datasets, supported writing, analysis, and dataset preparation for Hugging Face.
- **Xinyi (Alexis) Wu** Core contributor. Added several datasets, testing, and supported automatic metadata collection.
- **Enrico Shippole** Core contributor. Led final dataset preparation for Hugging Face upload and testing.
- **Kurt Bollacker** Advisor on project design and framing.
- **Tongshuang Wu** Advisor, particularly on data analysis and visualizations. Supported writing and Data Provenance Explorer design.
- **Luis Villa** Advisor on data copyright and licensing, and supporting writing in the legal discussion section.
- **Sandy Pentland** Advisor on general project design and framing.
- **Sara Hooker** Advisor on general project design and framing, as well as supporting writing, analysis, and directing experiments.

B Exact Licenses and Citations

See Table 5 for a summary of the Data Provenance Collection licenses and citations. More comprehensive details are available at <https://github.com/Data-Provenance-Initiative/Data-Provenance-Collection>.

C Details on Collecting Data Provenance

This data was collected with a mix of manual and automated techniques, leveraging dataset aggregators like GitHub, Hugging Face and Semantic Scholar. Annotating and verifying license information, in particular, required a carefully guided manual workflow, designed with legal practitioners (see Section 2.2). 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.

Automated Annotation Methods Based on the manually retrieved pages, we automatically extract Licenses from HuggingFace configurations and GitHub pages. We leverage the Semantic Scholar public API (Kinney et al., 2023) to retrieve the released date and current citation counts associated to academic publications. Additionally, we compute a series of other helpful, but often overlooked data properties such as text metrics (the min/mean/max for input and target lengths), and dialog 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 (Petrov et al., 2023).

API Annotation Methods While Task Categories have become the established measurement of data diversity in recent instruction tuning work (Sanh et al., 2021; Wang et al., 2022a), 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 10 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 4000 characters chunks and prompt the API to return a json list of any mentions of the dataset source, e.g. from scraping, synthetic or manual generation.

Collection	Cite	Licenses
Airoboros	Durbin (2023)	CC BY-NC 4.0
Alpaca	Taori et al. (2023)	CC BY-NC 4.0
Anthropic HH	Bai et al. (2022); Ganguli et al. (2022)	MIT License
BaizeChat	Xu et al. (2023b)	CC BY-NC 4.0
BookSum	Kryściński et al. (2022)	Academic Only
CamelAI Sci.	Li et al. (2023a)	CC BY-NC 4.0
CoT Coll.	Kim et al. (2023)	Non Commercial
Code Alpaca	–	Unspecified
CommitPackFT	Muennighoff et al. (2023a)	Various
Dolly 15k	Conover et al. (2023)	CC BY-SA 3.0
Evol-Instr.	Xu et al. (2023a)	Academic Only
Flan Collection	Longpre et al. (2023a)	Various
GPT-4-Alpaca	Peng et al. (2023)	CC BY-NC 4.0
GPT4AllJ	Anand et al. (2023)	Various
GPTeacher	–	Unspecified
Gorilla	Patil et al. (2023)	Apache License 2.0
HC3	Guo et al. (2023)	Various
Joke Expl.	–	MIT License
LAION OIG	Nguyen et al. (2023)	Various
LIMA	Zhou et al. (2023)	CC BY-NC-SA 4.0
Longform	Köksal et al. (2023)	CC BY-SA 3.0, Unspecified, CC BY-SA 4.0
OpAsst OctoPack	Muennighoff et al. (2023a)	CC BY 4.0
OpenAI Summ.	Stiennon et al. (2020)	CC BY 4.0
OpenAssistant	Köpf et al. (2023)	CC BY 4.0
OpenOrca	Mukherjee et al. (2023)	Various
SHP	Ethayarajh et al. (2023)	Unspecified
Self-Instruct	Wang et al. (2022a)	Apache License 2.0
ShareGPT	Vercel (2023)	Unspecified
StackExchange	–	Unspecified
StarCoder	Li et al. (2023b)	BigScience OpenRAIL-M
Tasksource Ins.	Sileo (2023)	Various
Tasksource ST	Weston et al. (2015)	Various
TinyStories	Eldan and Li (2023)	CDLA Sharing 1.0
Tool-Llama	Qin et al. (2023)	CC BY-NC 4.0
UltraChat	Ding et al. (2023)	CC BY-NC 4.0
Unnatural Instr.	Honovich et al. (2022)	MIT License
WebGPT	Nakano et al. (2021)	Apache License 2.0, CC BY-SA 4.0
xP3x	Muennighoff et al. (2022)	Various

Table 5: **Licenses and citations** for the dataset collections presented in this paper. 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.