BRIDGING THE DATA PROVENANCE GAP ACROSS TEXT, SPEECH, AND VIDEO

Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Manan Dey, Mohammed Hamdy, Nayan Saxena, Ahmad Mustafa Anis, Emad A. Alghamdi, Vu Minh Chien, Naana Obeng-Marnu, Da Yin, Kun Qian, Yizhi Li, Minnie Liang, An Dinh, Shrestha Mohanty, Deividas Mataciunas, Tobin South, Jianguo Zhang, Ariel N. Lee, Campbell S. Lund, Christopher Klamm, Damien Sileo, Diganta Misra, Enrico Shippole, Kevin Klyman, Lester JV Miranda, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Vipul Gupta, Vivek Sharma, Xuhui Zhou, Caiming Xiong, Luis Villa, Stella Biderman, Alex Pentland, Sara Hooker, Jad Kabbara

The Data Provenance Initiative

ABSTRACT

Progress in AI is driven largely by the scale and quality of training data. Despite this, there is a deficit of empirical analysis examining the attributes of well-established datasets beyond text. In this work we conduct the largest and first-of-its-kind longitudinal audit across modalities—popular text, speech, and video datasets—from their detailed sourcing trends and use restrictions to their geographical and linguistic representation. Our manual analysis covers nearly 4000 public datasets between 1990-2024, spanning 608 languages, 798 sources, 659 organizations, and 67 countries. We find that multimodal machine learning applications have overwhelmingly turned to web-crawled, synthetic, and social media platforms, such as YouTube, for their training sets, eclipsing all other sources since 2019. Secondly, tracing the chain of dataset derivations we find that while less than 33% of datasets are restrictively licensed, over 80% of the source content in widely-used text, speech, and video datasets, carry non-commercial restrictions. Finally, counter to the rising number of languages and geographies represented in public AI training datasets, our audit demonstrates measures of *relative* geographical and multilingual representation have failed to significantly improve their coverage since 2013. We believe the breadth of our audit enables us to empirically examine trends in data sourcing, restrictions, and Western-centricity at an ecosystem-level, and that visibility into these questions are essential to progress in responsible AI. As a contribution to ongoing improvements in dataset transparency and responsible use, we release our entire multimodal audit, allowing practitioners to trace data provenance across text, speech, and video.

1 INTRODUCTION

The capabilities and flaws of multimodal foundation models are often directly attributable to their training data (Carlini et al., 2023a; Rando et al., 2022; Carlini et al., 2023b; Parmar et al., 2024; Liu et al., 2023b;a; Dai et al., 2024). While the importance of *data measurement* has been widely established by prior work (Gadre et al., 2024), so has a prevailing absence of data documentation (Gebru et al., 2021; Bender & Friedman, 2018), transparency (Bommasani et al., 2023), and detailed understanding (Dodge et al., 2021; Bandy & Vincent, 2021; Sambasivan et al., 2021)—especially for modalities other than text. A lack of thorough data analysis has led to significant challenges, including privacy issues (Subramani et al., 2023), retracting datasets with harmful content (Birhane et al., 2021; David, 2023), adversarially bypassing safety filters (Rando et al., 2022), facial recognition bias with respect to gender and skin type (Buolamwini & Gebru, 2018a), gender bias in hiring (Chang, 2023), benchmark contamination from overlapping train and test sets (Lee et al., 2023a), and challenges in copyright (Henderson et al., 2023). Understanding data provenance can aid mitigation attempts to

	Datasets		Sources		Creator Orgs		Languages		Tasks	Licenses
	#	Size	#	Domains	#	Countries	#	Families		
Text	3717	2.1T	713	23	534	60	502	21	395	50
Speech	95	775k	51	16	124	29	260	36	18	19
Video	104	1.13M	44	24	101	23	-	-	33	11
Total	3916	-	798	83	659	67	608	37	443	55

Table 1: We quantify the breadth of our audit, including the total number of datasets (#), their size in tokens or hours, the sources, domains, creator organizations, countries, languages, tasks, and licenses. **In aggregate, we audited 3916 datasets from 659 organizations in 67 countries, spanning 2.1T tokens, and 1.9M hours. We cataloged nearly 798 unique sources, 443 tasks, and 55 licenses.**

reduce model bias and toxicity (Welbl et al., 2021; Pozzobon et al., 2023) address representation in data (Xu et al., 2021), contamination (Elazar et al., 2023), and quality (Kreutzer et al., 2022; Marion et al., 2023), as well as practical challenges with identifying copyright-free and permissively licensed sets (Min et al., 2023).

Despite the urgent need for the provenance and characteristics of widely used datasets, the majority of attention to date has centered on text datasets (Elazar et al., 2023; Longpre et al., 2024b), or a single feature such as prevalence of hate content or dataset (Dodge et al., 2021; Birhane et al., 2021). In contrast, in this work, we will critically examine several provenance features of data *across* text, speech, and video. We conduct the largest and most comprehensive multimodal audit of AI data, to date, reviewing nearly 4000 datasets between 1990-2024, covering 443 unique tasks, 608 languages, derived from 798 original sources, and constructed by 659 organizations, spanning 67 countries, over 1T tokens of text, and 1.9M hours of speech and video content (see Table 1).

There is an unprecedented acceleration in the development of multimodal AI systems, making all the more urgent an understanding of the datasets that underpin these breakthroughs. Our extensive collection of features from unstructured academic papers, websites, and repositories enables us to provide empirical grounding to an ambitious set of research questions surrounding data sourcing trends, intended licenses, and geographical and linguistic representation. Our key findings include:

1. **Multimodal data is increasingly sourced from the web, social media platforms, or synthetically generated;** rather than more curated sources such as movies, audiobooks or manually collected. These sources comprise the vast majority of text tokens, as well as speech and video hours in public data. However, while social media platforms provide data scale, heterogeneity and freshness by nature, they are also particularly prone to anti-crawling, copyright, privacy, and factuality concerns.
2. **Whereas only 25% of text, speech, and video datasets have non-commercial licenses, over 80% of content from each modality carries undocumented restrictions in the dataset's sources.** Dataset licenses are inconsistent with their source's restrictions for over 55% of content. Our audit provides the tools for multimodal developers to identify dataset restrictions, and apply their own standards.
3. **Geographical and linguistic representation have not improved for a decade, across the data ecosystem.** While the amount of data from under-represented creators and languages increases each year, to over 600 languages and 60 countries in 2024, their *relative representation* remains consistently western-centric, with no significant improvements from > 0.7 Gini coefficients. While Africa and South America organizations account for $< 0.2\%$ of all modality content, North America or European organizations span 93% of text tokens and 60%+ hours of speech and video.

Our work provides critical insights into the landscape of available multimodal data. We release the entire audit, collected data, and analysis tools, which we believe will bring immense value for data creators, developers, and researchers interested in promoting the responsible development of AI systems and analysis of the AI data ecosystem.

2 Methodology

While many prior works have surveyed the dataset ecosystem (Albalak et al., 2024; Liu et al., 2024b; Malik et al., 2021; Prabhavalkar et al., 2023; Li et al., 2019), few empirically examine data corpora at

scale, and those that do focus present a more narrow focus around a specific feature like geographic bias or hate content(Birhane et al., 2023; McMillan-Major et al., 2022a; Shankar et al., 2017) or single modality (Dodge et al., 2021; Caswell et al., 2021; Elazar et al., 2023; Longpre et al., 2024b). The goal of this work is to provide an empirical, ecosystem-level, and multimodal analysis of widely used training datasets (Cen et al., 2023). Our audit focuses on text, speech, and video, as prominent data modalities behind modern multimodal systems, such as Sora, Whisper, Gemini, GPT-4o, and others (Brooks et al., 2024; Zheng et al., 2024; Radford et al., 2023; Peng et al., 2023; Team et al., 2023; OpenAI, 2024). Since training data for modalities can often be independent, multimodal models tend to interleave training batches with different combinations of one or two modalities (Aghajanyan et al., 2023). As such, we focus our analysis on datasets that represent one or a pair of these modalities.

Annotation Features & Methodology In particular, we analyze data trends for the state of data permissions (licenses and terms), sourcing (the web, human annotation, and synthetic generation), and representation (of tasks, organizations, languages, and countries). We adopt Longpre et al. (2024b)'s methodology, including the license annotation taxonomy and process, to manually audit these features precisely and rigorously. We go beyond prior work, which considers dataset licenses, by extending the taxonomy to consider the terms of use of the sources of the dataset, either from models used to generate synthetic data (e.g. OpenAI's non-compete clause[1] or Meta's acceptable use policy for Llama 3.1[2]), or the source's policy on content restrictions, which can be conveyed in the form of a license, terms of use, or content policy on a website (Klyman, 2024). For each dataset, the source terms are annotated as Unrestricted, Unspecified, Source Closed or Model Closed, as defined in Table 2. For Figure 2 we combine Source Closed and Model Closed into *Restricted*.

As with prior work (Longpre et al., 2024b;c), we engage domain experts for these annotation tasks—AI researchers whose work pertains to the modality and topic. Because many datasets are iteratively re-packaged before they appear in their final form and often shared on popular dataset marketplaces like HuggingFace, Papers with Code or Github, prior work has found that relevant licensing terms or sourcing information for AI training data is frequently omitted (Longpre et al., 2024b). To ensure we collect this information, we require a full trace of metadata back to their original sources (sometimes a chain of github repositories, websites, or academic papers). This search can be onerous, especially for terms and licenses, but ensures rigor in the results. Table 1 enumerates the full statistics of our audit. All annotations and analysis code will be made publicly available on release.

Scope & Dataset Selection For each modality, we define the scope of the audit (detailed separately below), then aggregate resources to distill a list of relevant datasets. The scope is focused on (a) publicly available datasets, (b) widely used tasks in the context of general-purpose model development, and (c) relevance to generative tasks. However, we do consider classification-based datasets in text, speech, and video that can and are frequently re-purposed for generative uses (e.g. instruction tuning). Within the defined audit scope, we use a mix of the HuggingFace Datasets platform, survey papers, survey repositories, workshop proceedings, and expert review to accumulate relevant datasets. More detail about the dataset selection and collection process is given for each modality below. Each modality requires its own independent process, by virtue of their community dataset ecosystems being unique (discussed in Section 4). Note that text has a wider heterogeneity of published publicly available datasets than speech or video. Typically those datasets have been aggregated into large, standardized text-to-text collections, and as such we trace both these *Text (Collections)* and their constituent *Text (Datasets)*. All datasets are described, linked, and attributed in Appendix D.

2.1 TEXT

Scope We focus on providing an extensive audit for *post-training* datasets, used in training language models. We include single and multi-turn formats, encompassing both datasets typically used for instruction finetuning (SFT) and preference alignment Rafailov et al. (2023). This scope reflects the prominent role of general-purpose language models, which benefit from multi-task training on heterogeneous collections that span a variety of linguistic, reasoning, and knowledge intensive tasks like question answering, coding, tool use, translation, and classification (Wei et al., 2021; Ouyang et al., 2022).

Dataset Selection We expand the study conducted by the Data Provenance Collection (Longpre et al., 2024b), from 44 dataset collections (of 1858 supervised text datasets) to a superset of 108

[1] OpenAI Terms of Use

[2] Llama 3.1 Acceptable Use Policy

collections of 3717 datasets, prioritizing recent, popular publicly available HuggingFace Datasets introduced between 2022 and April 2024. Our collection sourced popular datasets from recent survey papers (Albalak et al., 2024; Liu et al., 2024b) and tools (Longpre et al., 2024a). We additionally reviewed HuggingFace Datasets' most downloaded datasets every month, from April to July 2024, under the Natural Language Processing category, as well as the SFT/DPO datasets associated with popular open model releases. We also drew from major multilingual data repositories, including the SEACrowd Catalogue (Lovenia et al., 2024), the Masader Arabic Data Catalogue (Alyafeai et al., 2022), AI4Bharat (Kunchukuttan et al., 2020), and the Aya Collection (Singh et al., 2024). Lastly, our list of datasets was reviewed and supplemented by language model experts to fill in notable omissions. In total, we trace the provenance and features of 3713 text datasets from 108 collections, covering 395 popular tasks, spanning from 1994 to 2024.

2.2 SPEECH

Scope We audit speech datasets for which automatic speech recognition (ASR) was noted as a primary task. We focus on ASR datasets because: (1) ASR is fundamental to many speech technologies, including dictation tools, voice assistants, and chatbots (Aksënova et al., 2021; Zhang et al., 2022); (2) large-scale speech datasets are typically designed for ASR (Li et al., 2023); (3) ASR data follows standardized formats, making comparisons easier (e.g., corpus of audio clips paired with text); and (4) ASR data can often be reused for other tasks like text to speech (TTS) (Ito & Johnson, 2017) or language identification (Ardila et al., 2020).

Dataset Selection To curate a representative sample of popular ASR datasets, we relied on a combination of survey repositories[3], and HuggingFace Datasets using the "Automatic Speech Recognition" and "Text-to-Speech" task tags. We expanded coverage to well-documented datasets on the OpenSLR[4] platform, even if they were newer or less widely used. We expect this might reflect datasets that could be adopted more widely in the future. Finally, we included datasets related to low-resource languages and other languages not well-covered by our initial searches. Speech recognition models are increasingly highly multilingual Babu et al. (2021); Radford et al. (2023); Pratap et al. (2024), and datasets serving different communities of builders and end-users around the world are a priority for making speech recognition technologies more inclusive. In total, we trace the provenance and features of 95 speech datasets, covering 18 popular ASR tasks, spanning from 1990 to 2024.

2.3 VIDEO

Scope Early video understanding models primarily focused on video classification, detection and action recognition, where short clips were categorized into predefined classes (Zheng et al., 2022; Zhu et al., 2020). More advanced tasks such as temporal action segmentation, video question answering, and video captioning were later introduced to build upon these foundational tasks (Moctezuma et al., 2022; Zhu et al., 2023). Recently, following the success in the field of image generation, video generation from text has become a new task that has shown promising results (Brooks et al., 2024; Zheng et al., 2024; Blattmann et al., 2023; Esser et al., 2023). Given the scarcity of datasets for text-to-video and the often undocumented sources of data used in recent video generation models (Mauran, 2024), we take a broader approach to our collection of video datasets. We focus on annotating popular video tasks and limit our scope to datasets corresponding to video tasks that are either published, highly cited, or have 100+ downloads on HuggingFace. This approach is justified by three key factors: (1) the usefulness of video data to the research community stems from its collection and presentation in peer-reviewed work, (2) datasets can often be repurposed between different tasks, allowing for applicability to new tasks such as video generation from text, and (3) focusing on highly cited datasets ensures that datasets' quality and relevance has been validated by the research community.

Dataset Selection We include datasets tagged with "Video Classification", "Text-to-Video", and "Video-Text-to-Text" from HuggingFace Datasets. We augmented this with datasets tagged by "Video Understanding" or "Video Generation" in PapersWithCode, as well as datasets listed in a popular Github survey repository. We also consulted the proceedings of recent video workshops: the Large Scale Video Understanding and Egocentric Vision workshops. We separately consulted a committee of non-author video experts to supplement the list with relevant datasets published at CVPR, ICCV,

[3]The Speech Datasets Collection

[4]openslr.org: Open Speech and Language Resources. OpenSLR is a widely used platform in the speech community, dedicated to hosting resources for speech tasks.

ECCV, and IJCV. In total, we trace the provenance and features of 104 video datasets, covering 33 popular video tasks, spanning from 2009 to 2024.

3 RESULTS

We discuss three key results related to (1) the rising use of web, social media and synthetic sources, (2) inconsistent and opaque restrictions on data use, and (3) a lack of improvement in geographical or linguistic representation. Each of these findings holds across modalities, at the ecosystem level.

3.1 RISING USE OF WEB, SOCIAL MEDIA & SYNTHETIC DATA

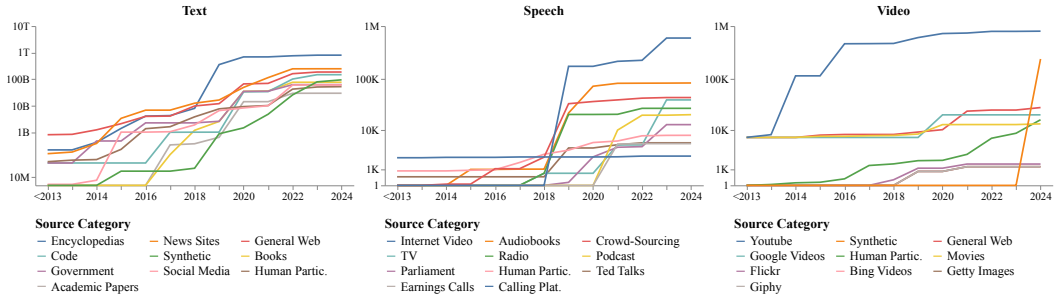


Figure 1: The cumulative size of data (log-scale tokens for text, hours for speech/video) from each source category, across modalities. The source categories in the legend are ordered by descending quantity. **Speech and video sources are increasingly dominated by internet videos and YouTube. Whereas text is predominantly web or encyclopedia-based (wiki) sources, synthetic text is rising in popularity.**

The need for scale, and heterogeneity have driven rising use of data from web-crawled, social media, and synthetic data sources. Developers have sought out ever larger and conveniently accessible sources of training data (Hoffmann et al., 2022; Henighan et al., 2020). While small, human-curated datasets are often sufficient and sometimes preferred due to higher quality, these sources often do not scale to present demands (Kaplan et al., 2020; Henighan et al., 2020). In Figure 1, we empirically measure the rising use of web crawling and social media (or "forum") websites that provide some of the most scalable and fresh content. While web-sourced data was always prominent, the balance of sources becomes much more skewed after 2018—note the use of the y-axis log scale. We find for Speech and Video that by far the most prominent source of data has become internet videos, and specifically YouTube. Nearly 1M hours each of Speech and Video data from this source far outstrips the next most common sources, which comprise less than 100K hours. For Speech, the primary data sources used to be Calling Platforms (pre-2017), content manually collected with Human Participation, and Audiobooks, but since 2018 internet videos have supplanted these other sources. For Video, since 2013, YouTube, synthetic, and general web data sources all constitute a significantly larger portion of data used in prominent video datasets, outstripping the use of Movies, Flickr, Getty, or human curated sources. Among text post-training datasets, we see a similar trend with general or news web-based sources, including encyclopedic sources (mainly Wikipedia), providing the majority of tokens over time. Encyclopedic sources alone now contribute over 1T tokens in total.

Synthetic data sources are rising the most rapidly. Within the video modality, the introduction of VidProm (Wang & Yang, 2024) in 2024, consisting of nearly 7M synthetically generated videos, offered a large shift in the video source distribution. Within the textual modality, from fig. 1, synthetic data represented <0.1% of the quantity of Web Encyclopedia data in 2020, but is now 10% its proportion in 2024, making up the 5th largest source of tokens. The top models used in generating datasets are mainly from OpenAI. The top 5 consist of ChatGPT, version unspecified (15.0% of synthetic datasets), GPT-4 (14.4%), BART (10.1%), GPT-3 (8.3%) and GPT-3.5-Turbo (4.9%). The average synthetic dataset also has notably longer turns (in tokens) than the average natural dataset: 1,756 tokens vs 1,065. The task distribution of textual synthetic datasets is shifted towards longer form, open-generation and creative tasks. For example, 88.1% of natural datasets contain classification tasks, compared to only 66.3% of synthetic datasets. Natural data is also more likely to cover translation than synthetic data (72.4% of datasets vs only 22.9% of synthetic datasets).

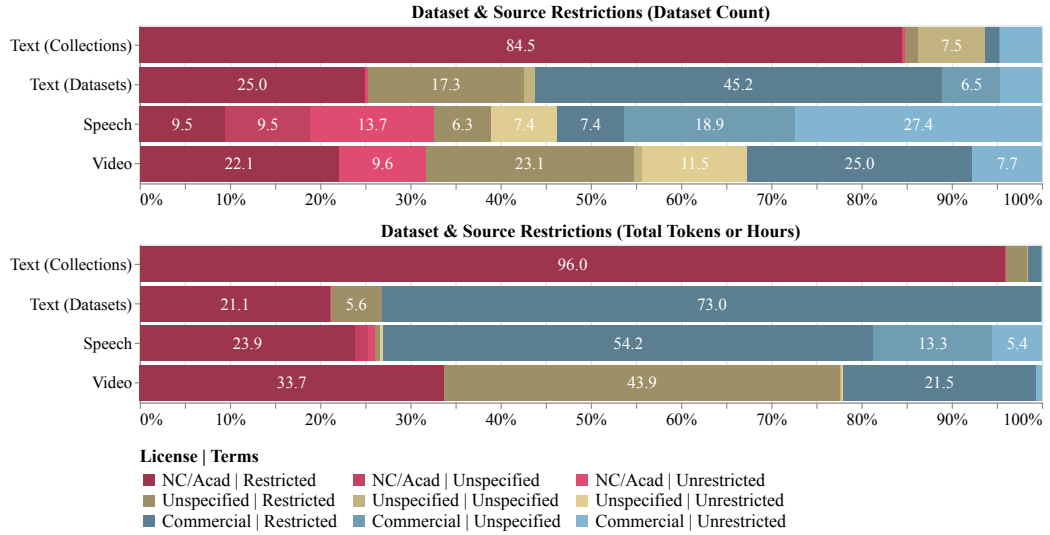


Figure 2: The distribution of restrictions from dataset *licenses* and their sources' *terms*. We break this down by the count of datasets (top), as well as total tokens or hours (bottom). Each license is categorized as Non-commercial/Academic (NC/Acad), Unspecified, or Commercially licensed. Each dataset may also have terms from the source: Restricted to non-commercial use, Unspecified restrictions, or Unrestricted. **Two main findings across modalities emerge: (1) Commercially licensed datasets represent a larger set of tokens and hours, relative to number of datasets; however, (2) the vast majority of those commercially licensed tokens/hours bare restrictions from their sources.** Tables 3 and 4 in the appendix provide detailed numbers.

3.2 Inconsistent Use Restrictions

In the United States, creators of a work automatically have a copyright interest that gives them exclusive rights to make copies and derivatives of the work (17 U.S.C. § 106). *Licenses* are legal documents through which the owners of a work express how others may use their work. By contrast, *Terms of Service* express a contract between a platform and its users to spell out how a platform and its content may be used (Robinson & Zhu, 2020). For simplicity, we use *"Licenses"* to refer to dataset restrictions, and *"Terms"* to refer to restrictions on the sources of datasets. There remain open questions about whether certain data licenses are enforceable, but these licenses signal the intention of data creators and therefore warrant consideration as the data creators may be best positioned to understand the sensitivities of the data (privacy, copyright, representation, etc.), and the most impacted by its downstream use (Morton-Park, 2023; Lee et al., 2023b; Mahari & Longpre, 2023; Mahari et al., 2023). The extent to which a practitioner adheres to dataset licenses or source terms remains an open question, and may depend on jurisdiction or the desired model's use cases (Lee et al., 2023b). *This work does not propose one standard for all developers.* For these reasons we restrict our treatment and discussion here to tracing the lineage and distribution of licenses and terms for a given modality.

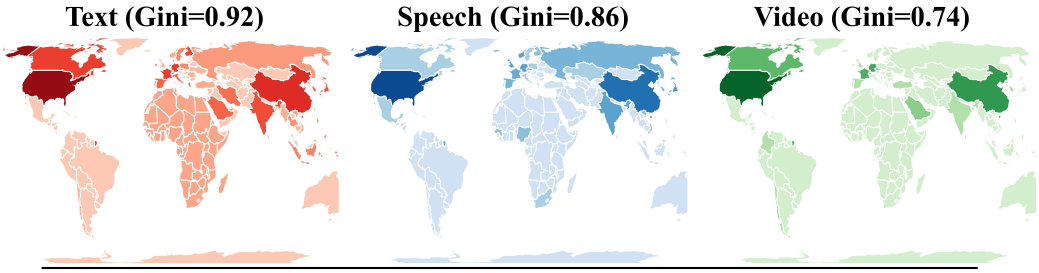
Data source terms are much more restrictive than the dataset's documented license restrictions. In Figure 2, we find only 25%, 33%, and 32% of text/speech/video datasets are licensed non-commercially. This value is even lower if we consider the proportion of tokens or hours, with 21%, 26%, and 33% of text/speech/video quantities carrying license restrictions. However, a staggering 99.8%, 78%, and 99% of those quantities carry some form of non-commercial restriction on one of their sources. For text, these restrictions are frequently from being generated by OpenAI or other models with a non-compete clause, while for speech and videos this is often since the datasets are derived from web or social media sources.

Inconsistencies between dataset licenses and their source's restrictions pose challenges to practitioners. A large amount of datasets have permissive or unspecified licenses, but some set of their sources carry non-commercial restrictions. This inconsistency is measurable—representing 79% of tokens in text datasets, 55% of speech hours, and 65% of video hours. Additionally, 19%,

14%, and 36% of text, speech, and video datasets have no license or intended use documentation (from our audit of the datasets' documentation on Hugging Face Datasets, GitHub, and Papers with Code). A lack of centralized documentation around these restrictions means it can be misleading to developers who are attempting to source data according to their own legal standards for copyright and privacy. Furthermore, lack of documentation can hamper developers following best practices around data preparation and transparency (Gebru et al., 2021; Bommasani et al., 2023).

Large quantities of commercially licensed text datasets are locked in collections without clear information to separate them from restrictive datasets. In Figure 2 (top and bottom), we see the number of datasets and number of tokens *without* restrictions is significantly higher for Text (Datasets) than Text (Collections). Specifically, 60% more Datasets (or 75% more tokens) are commercially licensed, than for Collections. This demonstrates that many collections contain significant amounts of commercially licensed data. While our audit traces licenses for all datasets within a collection, most collections do not aggregate or expose this documentation. As a result, practitioners may be left without easy access to filter for the subsets appropriate for their sourcing standards.

3.3 Geographical & Linguistic Representation is Not Improving



	Africa	Asia	Europe	N America	Oceania	S America
			By Count			
Text	0.3	13.4	24.0	61.5	0.7	0.2
Speech	3.6	35.7	30.4	30.4	0.0	0.0
Video	0.0	25.2	24.4	48.0	0.8	1.6
			By Tokens or Hours			
Text	0.0	6.1	55.4	38.4	0.1	0.0
Speech	0.1	38.8	18.8	42.4	0.0	0.0
Video	0.0	23.1	22.0	38.2	16.7	0.1

Figure 3: The geographical distribution of countries (world maps) and continents (table) represented by dataset creators. **Despite some differences in European, Russian, and Middle Eastern representation, creators are heavily concentrated in the US, China, and Western Europe, with little to no representation in South America or Africa, across modalities.** The current Gini coefficient for (Text, Speech, Video) = (0.92, 0.86, 0.74), where higher values indicate more concentration.

The importance and progress of representation in AI training data. Diversity and representation in training datasets, and among their creators, are widely acknowledged as essential to building AI models that are less biased, more useful, and more equitable (Joshi et al., 2020; Singh et al., 2024; Üstün et al., 2024; Adelani et al., 2021; 2024; Aakanksha et al., 2024; McMillan-Major et al., 2022b; Porgali et al., 2023; Monfort et al., 2019; Sigurdsson et al., 2016). Prior work has measured the diversity of languages in data along with cultural, ideological, and geographical imbalances (Faisal et al., 2022; Shankar et al., 2017; McMillan-Major et al., 2022a; De Vries et al., 2019; Mahadev & Chakravarti, 2021). These studies have exposed significant flaws, often in the form of bias and discrimination, stemming directly from poor representation in data (Buolamwini & Gebru, 2018b; Birhane et al., 2021). As this problem has now been widely acknowledged for decades, recent efforts have foregrounded sourcing data multilingually and multi-culturally, from native speakers and creators (e.g. ROOTS (Laurençon et al., 2022), the Aya Dataset (Singh et al., 2024), the SEACrowd Catalogue (Lovenia et al., 2024), the Masader Catalogue (Alyafeai et al., 2022), Common Voice

(Ardila et al., 2019), Causal Conversations V2 (Porgali et al., 2023) or Moments in Time (Monfort et al., 2019)).

Measuring geographical and linguistic representation. Naturally, we aim to use our audit to measure the progress of these efforts on geographical and linguistic representation in the AI ecosystem. We measure the progress of two forms of representation: (1) language diversity of text and speech data, and (2) geographical diversity of the creators, in all three modalities. For languages, we use the ISO 639-1 and 639-3 language codes and categories of language families from Glottolog 5.0.[5] In Figure 4(a, c) we display the cumulative sum of unique languages and countries present across all audited datasets, at each time period since 2013. While these measurements illustrate the absolute rise in diversity, we also hope to measure the relative dispersion, or equality of languages and countries in the distribution. In Figure 4(b, d), we use the Gini Index (Wilson, 1914; Atkinson et al., 1970), a traditional measure of statistical dispersion, frequently used to quantify inequality. This allows us to understand if the distributions of languages and creators are more representative of the international community over the last decade, or equally concentrated despite apparent efforts at the margins.

Inequality in geographical representation remains very high, with few organizations creating datasets from the Global South. For every dataset, our audit recorded the organizational affiliations of each creator of the dataset.[6] These organizations were then manually mapped to the country in which they are headquartered. Occasionally, organizations like BigScience, BigCode, or Masakhane have international or continental representation, and were counted as such. In Figure 3, we measure the current state of diversity among these creator organizations—where a Gini coefficient of 1 indicates highest concentration, and lower values more broad representation. Without taking up the normative question of what a truly "fair" score would be, these values provide useful comparisons across modalities and over time. We find that Text dataset developers are particularly homogeneous, with a Gini-coefficient of 0.92; followed by Speech, at 0.86 and Video at 0.74, which remain high, but are meaningfully less concentrated. Figure 3 also illustrates that even this limited diversity is still concentrated in North America, Europe, East Asia, and less so in the Global South.

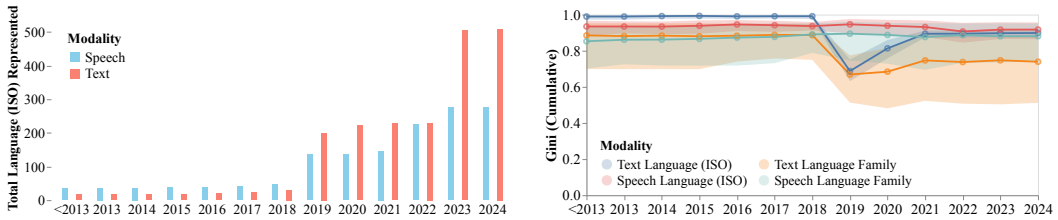
In Figure 3, we also compare the distribution of datasets, and of tokens or hours by continent. Dataset creators affiliated with African or South American organizations account for fewer than 0.2% of all tokens or hours, in each modality. In contrast, Asian affiliated organizations represent large proportions of the data, particularly for speech (39% of hours, attributed predominantly to YODAS (Li et al., 2023)). Much of this driven by Chinese, Indian, Russian, and Saudi Arabian creators. Most prominently, the combination of North American and European datasets comprises 93% of text tokens, 61% of speech hours, and 60% of video hours.

Geographical representation has not significantly improved for over a decade. In Figure 4(c), we measure the total unique number of countries represented across all dataset creator organizations. While individual creators will have varying ethnic and national affiliation, we treat this as an estimate for the influence of each locale in dataset development. We find that while the number of represented countries has risen steadily each year, for each modality, this represents only an illusion of progress. Empirically, the Gini coefficient for each modality has not significantly changed since the start of the period we examine in 2013. Geographic diversity has increased only among Video datasets, and these increases are not significant at the $p = 0.05$ level. Text and Speech geographical representations appear to remain stable over the last decade of AI development.

Multilingual representation has not improved by most measures. Similar to geographical representation, we measure the cumulative number of ISO 639-1 languages and language families over time, as well as the per-modality Gini-coefficient. Figure 4(a) shows significant increases in the number of languages available for speech and text, especially in 2019, and 2023, with the introduction of large sets like Flores (Goyal et al., 2022), xP3x (Muennighoff et al., 2023), Common Voice (Ardila et al., 2019), and the Aya Collection (Singh et al., 2024). However, once again, when measuring the cumulative dispersion of these datasets in Figure 4(b), only Text language families demonstrate any improvement from pre-2013 to the present. Improvements in the Gini coefficient appear to be largely driven by individual large-scale projects like xP3x and Common Voice, both introduced in 2019.

[5]We use top level Glottolog families.

[6]A dataset creator, following (Longpre et al., 2024b), is defined as an organization associated with the release of the dataset as created for machine learning—not any of the upstream sources. More details in Appendix D.



(a) Total Language Representation

(b) Gini-coefficient for Language Representation

(c) Total Geographical Representation

(d) Gini-coefficient for Geographical Representation

Figure 4: The cumulative totals (left) of languages and countries represented in the data over time, and the 95% confidence intervals of the gini-coefficients over time (right) to measure the representativeness of these variables. Gini-coefficients are a measure of statistical dispersion, frequently used to quantify inequality. A Gini coefficient of 1 indicates highest concentration, and lower values more broad representation. **While the number of represented languages and geographies continue to rise (left), the equality of their distribution has in most cases, not significantly changed.**

Subsequently, newer datasets remain predominantly monolingual, causing measures of concentration in text languages, speech languages, and language families to remain consistently high.

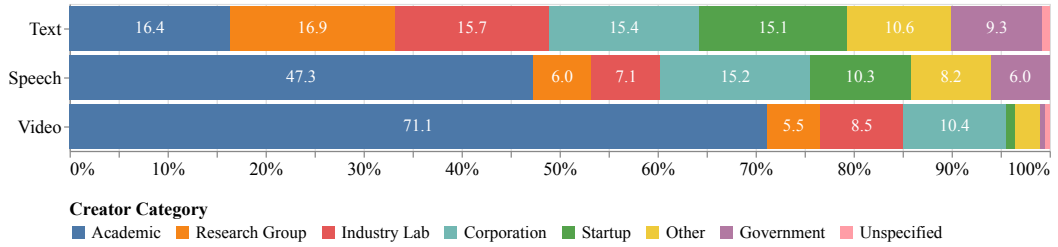


Figure 5: The distribution of creator organizations by modality. **Most public speech and video datasets are developed by academic organizations, whereas text datasets are developed by a wide mix of academia, non-profit or industry labs, as well as startups.**

Academia, research non-profits, and industry labs continue to drive public dataset development. As well as understanding the geographic associations of the organizations creating popular datasets, we manually categorize them into: Academic Organization (e.g., universities), Research Groups (e.g., non-profits such as BigScience, EleutherAI or AI2), Industry Labs (e.g., Cohere For AI, Google DeepMind), Corporations (e.g. Google, Meta), Startups (e.g., OpenAI, Anthropic), Governments, Unspecified (datasets where owner affiliation is not shared), or Other. When a dataset is released in collaboration between organizations, we record each organization. In Figure 5, we find that universities and other academic organizations account for 16%, 47%, and 71% of all recorded dataset releases, across Text, Speech, and Video respectively. Research groups, industry labs and even corporations are also significant contributors, especially for Text datasets, where ecosystem contributors are far more distributed. The significant role of academic organizations in Video and Speech may suggest that the risk profile of releasing Text datasets differs somewhat from Video and Speech datasets, which may have more distinct privacy concerns.

4 DISCUSSION

The rise of web-based, social media, and synthetic datasets may pose greater risks to privacy, copyright, and bias. Section 3.1 discusses the rise of web-based sources and particularly social media as primary sources for speech and video. Figure 1 shows these sources now exceed more traditional, curated sources such as movies, audiobooks, radio, TV, or content hand-crafted by human participants—by at least one order of magnitude. These websites made of mostly user-generated content are a natural choice, given that they scale in the quantity, freshness, and heterogeneity that is best suited to train general-purpose models (Longpre et al., 2023; Aghajanyan et al., 2023). However, prior work suggests that crowd-sourced, user-generated web content also introduces more challenges than curated content, particularly for privacy, copyright, bias, harm, and factuality.

Web-based and particularly user-generated content is disproportionately likely to include personally identifiable information (PII) Luccioni & Viviano (2021); Subramani et al. (2023); Elazar et al. (2023), and copyrighted content (Meese & Hagedorn, 2019; Lee et al., 2023b). These can be reproduced in the outputs of AI models (Carlini et al., 2022; Chen et al., 2023), creating privacy and copyright concerns (Zhang et al., 2023). Open datasets being used to train GPAI often attempt to filter—but frequently miss—PII and copyrighted data (Soldaini et al., 2024; Subramani et al., 2023) (although not all do (Penedo et al., 2023)). Social media, in particular, is also known to have bias, toxicity and factuality issues (Olteanu et al., 2019), which can manifest in trained models, even after alignment (Kotha et al., 2023). Lastly, while synthetic data can help reduce the prevalence of PII, copyright, or bias in data, it comes with its own challenges (Kurakin et al., 2023; Liu et al., 2024a).

Social Media websites have become one of the most prominent data sources, but their Terms often restrict crawling or commercial use. We find that 71% of Video data and 69% of Speech data is from YouTube which has become a prominent source of data, given its scale, freshness, and multimodality (containing videos, speech, images, and text) (Abu-El-Haija et al., 2016; Aytar et al., 2018; Chang et al., 2020; Uthus et al., 2023; Coats, 2023; Li et al., 2023). However, YouTube is a social media platform owned by Google and its Terms of Service[7] prohibit third parties from crawling YouTube. While content creators maintain their ownership rights in the material they upload to YouTube, the YouTube Terms of Service also grant Google a license to reproduce, modify, display, and use the content for purposes connected to YouTube's "business", which may include building machine learning models; even if the copyright holder has selected a permissive license, YouTube's Terms disallow external parties from crawling that data. Model developers such as Nvidia and OpenAI have been sued in the U.S. by content creators who allege that they unlawfully trained on YouTube videos (Cole, 2024; Skolnik, 2024). Large social media platforms and forums have also adopted restrictive terms in recent years, including Reddit and StackOverflow.[8] As these data sources become critical to scaling AI systems, access has been made exclusive, which may hamper academic, non-profit, or open source model development—to the extent that social media platforms can enforce their terms against third party developers.[9]

Ambiguous and poorly documented use restrictions may significantly inhibit model developers adhering to cautious legal and ethical data sourcing standards. In Section 3.2. we find that a significant amount of data carry non-commercial restrictions in their sources, rather than on the final dataset, which can contain no license or a permissive one. For text and video, these restrictions can equate to 99% of all tokens and hours. These inconsistencies are the result of datasets being iteratively re-packaged and re-licensed, without carrying on documentation (Longpre et al., 2024b). While not every developer will employ the same filtering standards, our work shows that the challenges to separate and identify appropriate datasets remain difficult across these modalities. Without continued audits and documentation, practitioners may be forced to forego large collections of partially viable data, hampering data scaling laws (Kaplan et al., 2020), or take on avoidable risk. We hope this released audit will provide greater tools for practitioners to apply their own standards, to make informed decisions on training data use.

The limitations of measures of geographical and linguistic representation. It is important to note that measures of geographical and linguistic representation are imperfect. We are limited by partial information about the developers' identities (including for privacy reasons), limited transparency

[7] YouTube Terms of Service.

[8] Reddit User Agreement and StackOverflow Terms of Service.

[9] We treat the enforceability of licenses and terms as an open legal question, beyond the scope of our work.

into how frequently these datasets are used, and the extent to which proprietary datasets may fill in representation gaps behind closed doors. Nonetheless, we believe the breadth and rigour of the audit make this the best available empirical measure of representation in *publicly* documented datasets. Further, we propose the goal of measuring representation in AI data as essential to understanding progress, or its absence, towards AI systems that fairly serve the broader community of users. Figure 3 and Figure 4 demonstrate that despite the absolute rise of geographical and linguistic representation, the relative western-centric concentration persists, across thousands of surveyed datasets. We release all audit materials for transparency and replicability, and for further use by the research community.

Conducting representative analyses of an ecosystem comes with assumptions. First, an ecosystem for AI is by nature, not centralized or organized. Widely used datasets for Text are often hosted on Hugging Face, but this is frequently not the case for Speech or Video. Similarly, while Text data undergoes frequent dataset re-packaging for general-purpose post-training, this is not true to the same extent for other modalities. As such, the scope and dataset selection process need to be designed for each modality, rather than a single, simple protocol, which inevitably will not accurately represent one modality at its ecosystem-level. Similarly, we chose a subset of modalities of interest to foundation model development (Brooks et al., 2024; Radford et al., 2023), but note there are many other left for future work (e.g., images, 3D representations, tabular, time series, graphs, and geospatial data).

ACKNOWLEDGMENTS

This research was conducted by the Data Provenance Initiative, a collective of independent and academic researchers volunteering their time to data transparency projects. The Data Provenance Initiative is supported by the Mozilla Data Futures Lab Infrastructure Fund.

REFERENCES

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. The multilingual alignment prism: Aligning global and local preferences to reduce harm, 2024. URL https://arxiv.org/abs/2406.18682.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. Irokobench: A new benchmark for african languages in the age of large language models, 2024. URL https://arxiv.org/abs/2406.03368.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. How might we create better benchmarks for speech recognition? In Kenneth Church, Mark Liberman, and Valia Kordoni (eds.), *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 22–34, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bppf-1.4. URL https://aclanthology.org/2021.bppf-1.4.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S Al-shaibani. Masader: Metadata sourcing for arabic text and speech data resources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6340–6351, 2022.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520.

Anthony B Atkinson et al. On the measurement of inequality. *Journal of economic theory*, 2(3): 244–263, 1970.

Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/35309226eb45ec366ca86a4329a2b7c3-Paper.pdf.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

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 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.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. Into the laions den: Investigating hate in multimodal datasets. *arXiv preprint arXiv:2311.03449*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,

Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018a. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018b. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. February 2022.

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*. OpenReview, 2023a.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023b. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Isaac Caswell, Julia Kreutzer, 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. *arXiv preprint arXiv:2103.12028*, 2021.

Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. AI Supply Chains, April 3 2023. URL http://dx.doi.org/10.2139/ssrn.4789403.

Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4283–4294. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/2cd4e8a2ce081c3d7c32c3cde4312ef7-Paper.pdf.

Xinyu Chang. Gender bias in hiring: An analysis of the impact of amazon's recruiting algorithm. *Advances in Economics, Management and Political Sciences*, 23:134–140, 09 2023. doi: 10.54254/2754-1169/23/20230367.

Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, June 2013. ISSN 1077-3142. doi: 10.1016/j.cviu.2013.01.013. URL http://dx.doi.org/10.1016/j.cviu.2013.01.013.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. Can language models be instructed to protect personal information? 2023.

Steven Coats. Dialect corpora from youtube. *Language and linguistics in a complex world*, 2023.

Samantha Cole. Nvidia sued for scraping youtube after 404 media investigation. *404 Media*, August 16 2024. URL https://www.404media.co/nvidia-sued-for-scraping-youtube-after-404-media-investigation/.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint*, 2024.

Emilia David. Ai image training dataset found to include child sexual abuse imagery. *The Verge*, December 2023. URL https://www.theverge.com/2023/12/20/24009418/generative-ai-image-laion-csam-google-stability-stanford. 7:57 AM PST.

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*, pp. 52–59, 2019.

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*, pp. 1286–1305, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. URL https://arxiv.org/abs/2302.03011.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. Dataset geography: Mapping language data to language users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3381–3411, 2022.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27092–27112. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/56332d41d55ad7ad8024aac625881be7-Paper-Datasets_and_Benchmarks.pdf.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

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.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020. URL https://arxiv.org/abs/2010.14701.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Keith Ito and Linda Johnson. The LJ Speech Dataset, 2017. URL https://keithito.com/LJ-Speech-Dataset.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science. *arXiv preprint arXiv:2207.07048*, 2022.

Kevin Klyman. Acceptable use policies for foundation models, 2024. URL https://arxiv.org/abs/2409.09041.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*, 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.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*, 2020.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. June 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 Frohberg, 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 (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31809–31826. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a.

Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin"bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023b.

Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312, 2019. doi: 10.1109/TETCI.2019.2892755.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M Dai. Best practices and lessons learned on synthetic data. April 2024a.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*, 2024b.

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, 2023.

Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, et al. The responsible foundation model development cheatsheet: A review of tools & resources. *arXiv preprint arXiv:2406.16746*, 2024a.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8): 975–987, August 2024b. doi: 10/gt8f5p.

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. *arXiv preprint arXiv:2407.14933*, 2024c.

Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Gero, Sandy Pentland, and Jad Kabbara. Data authenticity, consent, & provenance for ai are all broken: what will it take to fix them? *arXiv preprint arXiv:2404.12691*, 2024d.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P Kampman, et al. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *arXiv preprint arXiv:2406.10118*, 2024.

Alexandra Sasha Luccioni and Joseph D Viviano. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. May 2021.

Rohan Mahadev and Anindya Chakravarti. Understanding gender and racial disparities in image recognition models. *arXiv preprint arXiv:2107.09211*, 2021.

Robert Mahari and Shayne Longpre. Discit ergo est: Training data provenance and fair use. *Robert Mahari and Shayne Longpre, Discit ergo est: Training Data Provenance And Fair Use, Dynamics of Generative AI (ed. Thibault Schrepel & Volker Stocker), Network Law Review, Winter*, 2023.

Robert Mahari, Longpre Shayne, Lisette Donewald, Alan Polozov, Alex 'Sandy' Pentland, and Ari Lipsitz. Comment to US copyright office on data provenance and copyright, 2023.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457, 2021.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale, 2023. URL https://arxiv.org/abs/2309.04564.

Cecily Mauran. What was Sora trained on? Creatives demand answers. https://mashable.com/article/openai-sora-ai-video-generator-training-data, 2024. [Accessed 28-09-2024].

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*, 2022a.

Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, Nurulaqilla Khamis, Colin Leong, Maraim Masoud, Aitor Soroa, Pedro Ortiz Suarez, Zeerak Talat, Daniel van Strien, and Yacine Jernite. Documenting geographically and contextually diverse data sources: The bigscience catalogue of language data and resources, 2022b. URL https://arxiv.org/abs/2201.10066.

James Meese and Jennifer Hagedorn. Mundane content on social media: Creation, circulation, and the copyright problem. *Social Media+ Society*, 5(2):2056305119839190, 2019.

Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A Smith, and Luke Zettlemoyer. Silo language models: Isolating legal risk in a nonparametric datastore. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

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*, pp. 220–229, 2019.

Daniela Moctezuma, Tania Ramírez-delReal, Guillermo Ruiz, and Othón González-Chávez. Video captioning: a comparative review of where we are and which could be the route, 2022. URL https://arxiv.org/abs/2204.05976.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions, May 2021. URL http://arxiv.org/abs/2105.04489. arXiv:2105.04489 [cs, eess].

Frank Morton-Park. Licensed to learn: Mitigating copyright infringement liability of generative ai systems through contracts. *Notre Dame Journal on Emerging Technology*, 5:64, 2023.

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. Topics, authors, and institutions in large language model research: Trends from 17k arxiv papers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1223–1243, 2024.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, 2023.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019.

OpenAI. Hello gpt-4o: We're announcing gpt-4o, our new flagship model that can reason across audio, vision, and text in real time., 2024. URL https://openai.com/index/hello-gpt-4o/.

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.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Bo Liu, Aastha Jhunjhunwala, Zhilin Wang, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Data, data everywhere: A guide for pretraining dataset construction. *arXiv preprint 2407.06380*, 2024.

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. June 2023.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, et al. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset, 2023. URL https://arxiv.org/abs/2303.04838.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models, 2023. URL https://arxiv.org/abs/2310.07589.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv: arXiv:2204.06125, April 2022.

Francis McCann Ramirez, Luka Chkhetiani, Andrew Ehrenberg, Robert McHardy, Rami Botros, Yash Khare, Andrea Vanzo, Taufiquzzaman Peyash, Gabriel Oexle, Michael Liang, et al. Anatomy of industrial scale multilingual asr. *arXiv preprint arXiv:2404.09841*, 2024.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter, 2022. URL https://arxiv.org/abs/2210.04610.

Eric P Robinson and Yicheng Zhu. Beyond "i agree": Users' understanding of web site terms of service. *Social media+ society*, 6(1):2056305119897321, 2020.

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)*, pp. 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.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. Include: Evaluating multilingual language understanding with regional knowledge, 2024. URL https://arxiv.org/abs/2411.19799.

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 *CHI*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

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.

Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, July 2023. ISSN 1557-7341. doi: 10.1145/3577925. URL http://dx.doi.org/10.1145/3577925.

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.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding, 2016. URL https://arxiv.org/abs/1604.01753.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. arXiv: arXiv:2209.14792, September 2022. URL http://arxiv.org/abs/2209.14792.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

Sam Skolnik. Openai sued over using youtube videos without creators' consent. *Bloomberg Law*, August 5 2024. URL https://news.bloomberglaw.com/litigation/openai-sued-over-using-youtube-videos-without-creators-consent.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

Dave Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29029–29047. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5c61452daca5f0c260e683b317d13a3f-Paper-Datasets_and_Benchmarks.pdf.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *arXiv preprint arXiv:2403.06098*, 2024.

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*, pp. 2447–2469, 2021.

E. B. Wilson. Untitled review. *The American Economic Review*, 4(2):442–444, 1914. ISSN 00028282. URL http://www.jstor.org/stable/1804762.

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*, pp. 2390–2397, 2021.

Xinyu Yang, Weixin Liang, and James Zou. Navigating dataset documentations in ai: A large-scale analysis of dataset cards on hugging face, 2024. URL https://arxiv.org/abs/2401.13822.

Dawen Zhang, Boming Xia, Yue Liu, Xiwei Xu, Thong Hoang, Zhenchang Xing, Mark Staples, Qinghua Lu, and Liming Zhu. Tag your fish in the broken net: A responsible web framework for protecting online privacy and copyright. October 2023.

Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou, Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang, Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran, Tara N. Sainath, Francoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu, Ruoming Pang, and Yonghui Wu. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1519–1532, October 2022. ISSN 1941-0484.

doi: 10.1109/jstsp.2022.3182537. URL http://dx.doi.org/10.1109/JSTSP.2022.3182537.

Lu Zheng, Tongtong Zhou, Rongqi Jiang, and Yueping Peng. Survey of video object detection algorithms based on deep learning. In *Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '21, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450385053. doi: 10.1145/3508546.3508622. URL https://doi.org/10.1145/3508546.3508622.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL https://github.com/hpcaitech/Open-Sora.

Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review, 2023. URL https://arxiv.org/abs/2302.12552.

Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition, 2020. URL https://arxiv.org/abs/2012.06567.

A Extended Related Work

Progress in machine learning across modalities from speech (Radford et al., 2023) to vision (Dosovitskiy et al., 2021) to text (Brown et al., 2020; Wei et al., 2021) has benefited from advancements in large pre-training and fine-tuning corpora. The development of multimodal corpora has also been key to several recent advances, as with CLIP in the image/text domain Radford et al. (2021), CLAP for audio/text settings Elizalde et al. (2022), and a number of other models involving both text and images, audio or video (Radford et al., 2023; Ramirez et al., 2024; Singer et al., 2022; Ramesh et al., 2022).

The datasets powering these advances are not, however, always well-documented, despite the existence of standards and frameworks for recording and annotating dataset metadata that range from 'data statements' (Bender & Friedman, 2018) to 'datasheets for datasets' (Gebru et al., 2021) and others (Mitchell et al., 2019). The key problem is not a deficiency of any particular framework, but rather inconsistent adoption and fragmentation (Longpre et al., 2024d). Much prior work has argued for the need to document and audit these datasets (Rogers, 2021; Paullada et al., 2021), motivated by concerns from reproducibility (Kapoor & Narayanan, 2022) to interpretability (Longpre et al., 2023) to bias and fairness problems that may stem from problematic content in training data (Birhane et al., 2021).

There have been several attempts to carry out such audits, with prior work examining pretraining data (Longpre et al., 2024c), general web corpora (Gao et al., 2020; Dodge et al., 2021), instruction finetuning datasets (Longpre et al., 2024b), and the documentation fields of the HuggingFace Datasets platform in particular (Yang et al., 2024). For speech and vision, there has been less work, with many discussions of datasets in the aggregate occurring in survey papers (Schiappa et al., 2023; Chaquet et al., 2013), research aimed directly at improving model performance Gadre et al. (2023) or close examinations of questions like bias in small groups of datasets (Buolamwini & Gebru, 2018b; Romanou et al., 2024).

Prior work has also examined the identities, affiliations and national origin of paper authors (Movva et al., 2024) in AI, but an analogous look at the producers of datasets is lacking. We aim to carry out such analyses: replicating those for pretraining and text finetuning datasets in video and audio domains, and surveying provenance and legal status. Finally, there has also been significant recent attention to legal questions in the collection and use of AI training data (Sag, 2020; Henderson et al., 2023). The complex process involved in preparing these datasets (Lee et al., 2023b), and the ambiguous licensing of inputs, can make understanding the legal status of the final output quite difficult.

B Dataset Licenses & Terms

Detailed taxonomy We code the legal restrictions placed on use of datasets along two axes. First, we identify whether a dataset's license permits commercial use ("Commercial" in Table 3), only non-commercial / academic use ("NC / Acad"), or does not clearly specify what is permitted ("Unspecified"). The latter category includes datasets for which we were unable to locate a license. Datasets which are in the public domain and not subject to a license are counted as commercially usable. Second, we annotate the contractual or terms-of-use restrictions placed on dataset use by the source of each dataset. There are four levels, defined in Table 3. Note that the Model Closed status can only apply to datasets that are AI-generated, at least in part. Some datasets can carry both Model Closed and Source Closed status, but we count the Model Closed first for simplicity.

Detailed breakdown Tables 3 and 4 present crosstabs of these two dimensions, according to respectively the total amount of content and the number of datasets. The most notable finding, as discussed in the main text, is the frequency of clashing restriction status between licenses and terms. By amount of content, fully 73.0% of text content, 55.0% of speech content, and 21.6% of video content is subject to a license permitting commercial use but also to terms restrictions forbidding it, or the reverse. The absolute level of restrictions is also high, with < 0.1% of text content, 5.4% of speech content, and 0.6% of video content usable for commercial purposes under both licenses and terms.

LABEL	DEFINITION
MODEL CLOSED	A model used to generate part or all of the dataset prohibits using its outputs commercially, to develop a competing AI model, or in general.
SOURCE CLOSED	The source has a license or terms that prohibits use of the data, either commercially, from being crawled, to develop AI, or in general.
UNSPECIFIED	No information can be found relevant to restrictions, or lack thereof, for this source.
UNRESTRICTED	The source has a commercially permissive license, such as CC BY, or explicitly states the data is open for broad use.

Table 2: **The taxonomy used to determine use restrictions on each dataset source.** Each source in a dataset is examined and fit into one of these categories. The dataset Terms are then labelled according to the strictest terms across the sources, with Model Closed and Source Closed considered stricter than Unspecified which is in turn stricter than Unrestricted.

LICENSE / TERMS	RESTRICTED	UNSPECIFIED	UNRESTRICTED	TOTAL
Text Collections				
NC/ACAD	96.0	0.0	0.0	96.0
UNSPECIFIED	2.3	0.1	0.0	2.4
COMMERCIAL	1.5	0.0	0.0	1.6
TOTAL	99.8	0.1	0.1	
Text Datasets				
NC/ACAD	21.1	0.0	0.0	21.2
UNSPECIFIED	5.7	0.1	0.0	5.7
COMMERCIAL	73.0	0.0	0.0	73.1
TOTAL	99.8	0.1	0.1	
Speech Datasets				
NC/ACAD	23.9	1.4	0.8	26.2
UNSPECIFIED	0.5	0.0	0.4	0.9
COMMERCIAL	54.2	13.3	5.4	73.0
TOTAL	78.6	14.7	6.7	
Video Datasets				
NC/ACAD	33.7	0.0	0.1	33.8
UNSPECIFIED	43.9	0.1	0.1	44.1
COMMERCIAL	21.5	0.0	0.6	22.1
TOTAL	99.1	0.1	0.8	

Table 3: **A breakdown of the percentage of license and terms restrictions across datasets**, by total tokens or hours of content. The much higher frequency of restrictions at the collection level is because we consider a collection's license or terms status to be the most restrictive of those for its datasets. Note that percentages may not add to exactly 100% because of rounding.

C ADDITIONAL RESULTS

Figures 6 and 7 report the size distributions of the datasets. We measure size differently for different types of datasets: Text datasets are in tokens, and audio/video in hours of content. The lack of standard tokenization or preprocessing schemes for those modalities makes it simplest to report raw dataset size.

Notably, we find quite different size distributions by modality. The distribution of dataset sizes has the thickest right tail for text, followed by speech and then by video. Most video datasets are short in

License / Terms	Restricted	Unspecified	Unrestricted	Total
Text Collections				
NC/Acad	84.5	0.0	0.3	84.8
Unspecified	1.5	7.5	0.0	8.9
Commercial	1.5	0.2	4.5	6.3
Total	87.5	7.7	4.8	
Text Datasets				
NC/Acad	25.0	0.0	0.3	25.3
Unspecified	17.3	1.2	0.0	18.5
Commercial	45.2	6.5	4.5	56.2
Total	87.5	7.7	4.8	
Speech Datasets				
NC/Acad	9.5	9.5	13.7	32.6
Unspecified	6.3	0.0	7.4	13.7
Commercial	7.4	18.9	27.4	53.7
Total	23.2	28.4	48.4	
Video Datasets				
NC/Acad	22.1	0.0	9.6	31.7
Unspecified	23.1	1.0	11.5	35.6
Commercial	25.0	0.0	7.7	32.7
Total	70.2	1.0	28.8	

Table 4: **A breakdown of the percentage of license and terms restrictions** by dataset count. The much higher frequency of restrictions at the collection level is because we consider a collection's license or terms status to be the most restrictive of those for its datasets. Note that percentages may not add to exactly 100% because of rounding.

hour terms, with speech datasets tending to be somewhat longer and text datasets having a greater prevalence of both very small and very large datasets relative to the mean size.

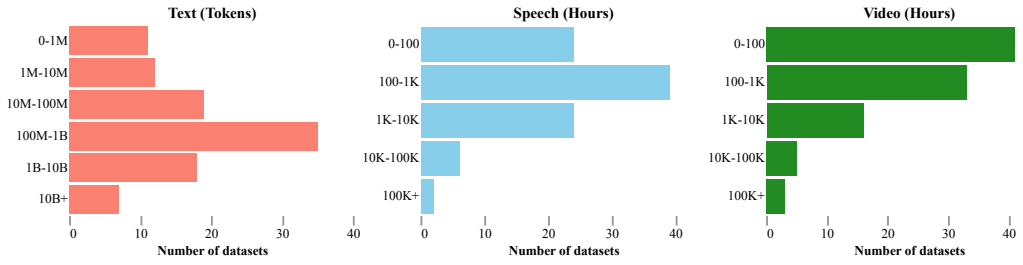


Figure 6: The distribution of dataset sizes for each modality. Most text data collections are between 100M-1B tokens. **Speech datasets average 100-1k hours, and video datasets are usually the smallest, commonly less than 100 hours.**

Dataset tasks, meanwhile, reflect traditional approaches and research programs for each modality. Classification is the most common task for both text and video, with the video community's long-standing interest in captioning also visible in its role as the second most common task for video datasets. Q&A occupies a similar role for text, though text datasets have a more balanced distribution over other, increasingly prominent tasks like generation and reasoning. Given our selection criteria, all datasets for speech are for ASR tasks, but other tasks like speaker identification and translation are also represented.

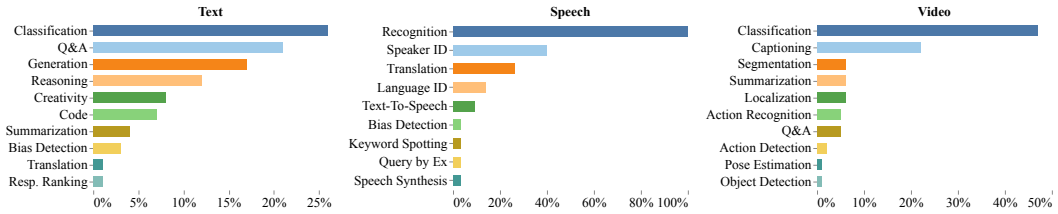


Figure 7: The task distribution of datasets, across modalities. Post-training text and video datasets are predominantly based on classification. For text, generation and reasoning are rising categories. All speech datasets are recognition-based, particularly for speaker, language, or in the process of translation.

D DATASETS

This section provides a detailed overview of the datasets we have collected and analyzed. Table 5 summarizes the text datasets, Table 6 the audio datasets, and Table 7 the video datasets. Each of these tables lists broad collections of data, sorted in chronological order, and provides information about their properties, sizes, sources and permissions. Each collection can include multiple datasets, and they generally reflect the ways dataset creators have grouped their datasets (such as in the same paper). Because of the large number of datasets, we provide detailed information about their licenses and original published papers, where applicable, in the supplementary Attribution Card.

Annotation Details: Text For post-training text datasets it is common to package many together as collections, such as Flan (Wei et al., 2021) or P3 (Sanh et al., 2021). This practice is not common to the same extent for speech or video datasets. For much of the text analysis, where possible, we chose to analyze statistics at the collection-level, since practitioners are more likely to adopt a collection for general-purpose post-training, than an individual dataset within the collection. Also, in dataset-level statistics, metadata for a single collection with many datasets can get repeated and overwhelm the statistics unfairly (e.g. the dataset aggregator/creator being repeated hundreds of times). Consequently, our collection-level analysis of the text modality is reflected in Figure 1, Figure 3, Figure 5, Figure 4, Figure 7, and Figure 6. However, for Figure 2 we draw the distinction between collection and dataset metrics, as practitioners may wish to unpack collections to extract only commercially licensed data. In that case a Collection inherits the most restrictive license and terms of its constituent datasets.

For annotating creator organizations, we follow prior work's instructions (Longpre et al., 2024b). For each dataset they record the affiliations listed on the academic paper or GitHub or HuggingFace object in which the dataset was released. This does not include the organizations who created or owned the sources from which the data was derived. For instance, the SQuAD dataset (**?**) would be associated with Stanford (the authors' affiliation), but not Wikipedia, which the data was partially derived from. For a dataset that has authors affiliated with multiple organizations, the dataset will be counted towards each organization.

Annotation Details: Speech In many cases, multiple versions of a dataset exist due to datasets being expanded or updated. In these scenarios, we used the release date from the initial version (since release dates for subsequent versions were not always clear), but used metadata from the most recently released version for which information was available to offer an overview of the current landscape of data. However, if the dataset versions could not be meaningfully aggregated (e.g. different licenses), or did not appear to be cumulatively designed (non-overlapping or otherwise semantically disjoint data), we maintained separate records. We kept only datasets for which ASR was noted as a primary task. For example, if a dataset was primarily intended for text-to-speech or speaker recognition, we did not keep it even if it could conceivably be repurposed for ASR. When computing hours, we excluded any hours without supervisory transcripts/scripts (unlabeled data), but kept hours with "weak supervision" (e.g. model-generated transcripts from speech audio). We recognize the difficulty in comprehensively covering all relevant datasets.

Annotation Details: Video In video, a single dataset can be re-purposed and annotated to address different tasks Monfort et al. (2019; 2021). We consider these as two different datasets even if they have the same video source since now they can be used for different computer vision tasks.

Table 5: **Alignment tuning (text) collections and properties**. Collection properties include numbers of datasets, tasks, languages, and text domains. The SOURCE column indicates whether a collection contains human-generated web text (🌐), language model outputs (🤖) or both (🌐🤖). The USE column indicates whether a collection includes data freely usable even for commercial purposes (🔵), data usable only for noncommercial purposes or academic research (🔴) and data whose license status is not specified precisely enough to allow us to determine commercial use permissions (🟡). Note that each collection may have different datasets with one, two, or all three of these statuses. Finally, the OAI column indicates collections which include OpenAI model generations. Datasets are sorted chronologically to highlight trends over time.

COLLECTION		PROPERTY COUNTS				TYPES	PERMISSIONS	
	YEAR	DATASETS	TASKS	LANGS	DOMAINS	SOURCE	USE	OAI
RiddleSense	2021	1	3	1	1	🌐	🔵	
MathInstr.	2023	1	3	1	1	🤖	🔵	✔
No Robots	2023	1	8	1	1	🌐	🔴	
Nectar	2023	1	1	1	2	🤖	🔵🔴	✔
MetaMathQA	2023	8	2	1	1	🤖	🔵	✔
MegaWika	2023	50	1	50	1	🤖	🔵	
MedInstr.	2023	1	1	1	1	🤖	🟡	✔
MathDial	2023	1	2	1	4	🤖	🔵	✔
PII-Masking-200k	2023	1	2	4	1	🌐	🔴	
Pure-Dove	2023	1	4	1	1	🤖	🔵	✔
LMSYS-Chat-1M	2023	1	9	5	1	🤖	🔵🔴	✔
PygmalionAI-PIPPA	2023	1	3	1	1	🤖	🔵	
HelpSteer	2023	1	5	1	1	🌐	🔵	
SeaBench	2023	9	4	9	5	🤖	🔵	
Open Asst. v2	2023	19	4	19	1	🌐	🔵	
Feedback Coll.	2023	1	2	1	1	🤖	🔵	✔
Glaive Code Asst.	2023	1	2	2	1	🤖	🔵	
EverythingLM	2023	1	8	2	1	🤖	🔵	✔
Bactrian-X	2023	6	4	6	1	🤖	🔵🔴	✔
COBRA Frames	2023	1	1	1	2	🤖	🔵	✔
UltraFeedback Argilla	2023	9	16	1	20	🌐🤖	🔵🟡🔴	✔
ExpertQA	2023	1	3	1	1	🤖	🔵	✔
ChatDoctor	2023	3	1	1	2	🌐🤖	🟡	✔
Capybara	2023	11	17	2	1	🤖	🔵🟡🔴	✔
UltraChat-200k	2023	1	7	1	2	🤖	🔴	✔
CollectiveCognition	2023	1	6	1	1	🤖	🔵	✔
Thai Gen AI	2023	9	11	1	1	🤖	🔵🔴	✔
Deita 10K	2023	2	11	1	3	🤖	🔵🔴	✔

Continued on next page

Table 5: **Alignment tuning (text) collections and properties**.

Collection	Year	Property Counts				Types	Permissions	
		Datasets	Tasks	Langs	Domains	Source	Use	OAI
SelFee	2023	1	5	1	1	🤖	🔵	✔
ChatbotArena	2023	1	4	1	1	🤖	🔵🔴	✔
OpenGPT Healthcare	2023	3	4	1	1	🤖	🔵🟡	✔
Orca-Math	2024	1	1	1	3	🤖	🔵🔴	✔
OpenMathInstr.-1	2024	2	3	1	3	🤖	🔵🟡	
WildChat	2024	2	7	10	1	🤖	🔵	✔
Magpie-Pro	2024	1	9	1	1	🤖	🔵	
10k Prompt Ranked	2024	1	13	1	4	🤖	🟡	✔
Synth.-GSM8K-Refl.	2024	1	3	1	1	🤖	🔵	
LongAlign-10k	2024	1	3	1	1	🤖	🔵🔴	✔
Llama2-MedTuned-Instr.	2024	1	4	1	1	🌐	🔴	
KIWI	2024	1	1	1	2	🤖	🔵	✔
Indic-Instr.	2024	8	7	2	3	🤖	🔵🟡🔴	✔
Gretel Text-to-SQL	2024	1	1	3	1	🤖	🔵	
Conifer	2024	1	8	1	2	🤖	🔵	✔
Cidar	2024	1	8	1	1	🤖	🔴	✔
Aya	2024	71	7	71	1	🌐	🔵	
Reasoning	2024	1	4	1	1	🤖	🔵	✔
AgentInstruct	Mult.	6	3	1	7	🌐🤖	🔵🟡	✔
InstAr	Mult.	24	13	1	9	🌐🤖	🔵🟡🔴	✔
Dynosaur	Mult.	1k	21	1	22	🌐🤖	🔵🟡🔴	✔
Medical Meadow	Mult.	8	2	1	3	🌐🤖	🔵🟡🔴	✔
Open-Platypus	Mult.	10	10	36	8	🌐🤖	🔵🟡🔴	✔
PMC-LLaMA Instr.	Mult.	7	1	1	2	🌐🤖	🔵🟡	✔
COIG	Mult.	18	13	2	22	🌐🤖	🔵🟡🔴	
DialogStudio	Mult.	83	3	5	3	🌐🤖	🔵🟡🔴	

Table 6: **Audio collections and properties**. Collection properties include numbers of audio hours (HR), speakers (SPKR), languages (Lang), creator institutions (Creat), tasks (Tasks), data sources (Src), and topics (Topics). The number of datasets is not listed because all collections include only one dataset, except for M2ASR which has four. The US column indicates datasets from or partly from the United States, the Ac column datasets created by academic institutions, and the Ind column datasets created by industry. Note that a dataset can have all of these, none of them, or any combination of them. The Use column indicates whether a collection includes data freely usable even for commercial purposes (🔵), data usable only for noncommercial purposes or academic research (🔴) and data whose license status is not specified precisely enough to allow us to determine commercial use permissions (🟡). Note that each collection may have different datasets with one, two, or all three of these statuses. Datasets are sorted chronologically to highlight trends over time.

Collection	Year	Property Counts							Category			Perm
		Hr	Spkr	Lang	Creat	Tasks	Src	Top	US	Ac	Ind	Use
TIMIT	1990	5	630	1	3	3	1	7	✔	✔	✔	🔴
Switchboard	1992	250	543	1	1	1	1	70	✔		✔	🔴

Continued on next page

Table 6: **Audio collections and properties**.

Collection	Year	Property Counts							Category			Perm
		Hr	Spkr	Lang	Creat	Tasks	Src	Top	US	Ac	Ind	Use
African Acc. French	2003	22	232	1	1	1	1	7	✔			
CSJ	2003	661	1k	1	1	1	1	2				
Fisher	2004	2k	12k	1	1	1	1	36	✔	✔		
CSLU 22 Langs.	2005	84	-	21	1	1	1	7	✔	✔		
AMI	2005	100	-	1	1	1	2	2		✔		
CSLU 1.2	2007	25	5k	1	1	1	1	1	✔	✔		
ALLSSTAR	2010	86	140	27	1	1	1	3	✔	✔		
TED-LIUM3	2012	452	2k	1	2	2	1	1		✔	✔	
NST Norwegian	2013	540	870	1	1	1	1	7				
NST Danish	2013	500	-	1	1	1	1	7				
NST Swedish	2013	300	-	1	1	1	1	7				
Vystadial	2014	56	-	2	1	1	2	3		✔		
THCHS-30	2015	35	40	1	1	1	1	1		✔		
LibriSpeech	2015	1k	2k	1	1	1	1	106	✔	✔		
THUYG-20	2015	20	371	1	2	2	1	3		✔		
VCTK	2016	44	110	1	1	1	1	1		✔		
Spoken Wikipedia	2016	1k	960	3	1	1	1	1		✔		
AISHELL-1	2017	520	400	1	2	2	2	11			✔	
LJSpeech	2017	24	1	1	1	1	1	1	✔			
ClarinPL	2017	56	317	1	1	1	2	7		✔		
AISHELL-2	2018	1k	2k	1	2	2	1	8			✔	
Regional Af. Am. Lang.	2018	159	222	1	1	1	1	8	✔	✔		
Crowd Sourced Speech	2018	1k	3k	5	1	1	1	1	✔		✔	
Zeroth-Korean	2018	96	181	1	1	1	1	7			✔	
RTVE	2018	691	-	1	1	1	1	7		✔		
OpenSTT	2019	20k	-	1	2	2	2	6		✔	✔	
MuST-C	2019	4k	2k	16	2	2	1	4		✔		
M-AILABS	2019	1k	-	8	1	1	1	33				
MAGICDATA	2019	755	1k	1	1	1	1	1			✔	
Common Voice 17	2019	31k	330k	124	3	3	1	1	✔	✔	✔	
CoNASE	2019	154k	-	1	1	1	1	6		✔		
Nigerian English	2019	6	-	1	1	1	1	7	✔		✔	
Norwegian Parl. Speech	2019	140	309	1	1	1	1	7				
120h Spanish Speech	2019	120	17	1	1	1	1	7				
DiDiSpeech	2020	800	6k	1	1	1	1	2			✔	
Czech Parliament	2020	444	212	1	1	1	1	7		✔		
CoVoST-2	2020	3k	78k	22	1	1	2	1	✔		✔	
KSC	2020	332	-	1	1	1	1	5		✔		
Basq., Cat. and Gal.	2020	34	132	3	1	1	1	2	✔		✔	
KsponSpeech	2020	969	2k	1	1	1	1	6				
Samromur	2020	145	8k	1	1	1	1	5		✔		
Multiling. LibriSpeech	2020	50k	6k	8	1	1	1	33	✔		✔	
MaSS	2020	160	-	8	1	1	1	1		✔		

Continued on next page

Table 6: **Audio collections and properties**.

Collection	Year	Property Counts							Category			Perm
		Hr	Spkr	Lang	Creat	Tasks	Src	Top	US	Ac	Ind	Use
FT SPEECH	2020	2k	434	1	2	2	1	2	✔	✔	✔	
Eng. Acc. in Brit. Isles	2020	31	120	1	1	1	1	4			✔	
Highland Puebla Nahuatl	2021	156	-	1	3	3	1	7	✔	✔		
QASR	2021	2k	11k	1	2	2	1	7	✔	✔	✔	
Multiling. TEDx	2021	765	-	9	3	3	1	7	✔	✔		
Minds14	2021	25	-	14	1	1	2	7			✔	
Golos	2021	1k	-	1	3	3	1	6		✔	✔	
MASC	2021	1k	14k	1	3	3	1	15		✔	✔	
LaboroTVSpeech	2021	2k	-	2	2	2	1	7		✔	✔	
KeSpeech	2021	2k	27k	2	1	1	1	1		✔		
JTUBESPEECH	2021	1k	-	2	4	4	1	7	✔	✔		
GigaSpeech	2021	10k	-	1	9	9	3	24	✔	✔	✔	
VoxPopuli	2021	2k	4k	16	1	1	1	1	✔		✔	
SPGISpeech	2021	5k	50k	1	4	4	1	2	✔	✔	✔	
West Afr. Radio	2021	142	-	10	2	2	1	3	✔	✔		
AISHELL-4	2021	120	61	1	4	4	2	6	✔	✔	✔	
West Afr. Virt. Asst.	2021	2	49	3	2	2	1	2	✔	✔		
MediaSpeech	2021	40	-	4	5	5	12	1		✔	✔	
People's Speech	2021	30k	-	1	7	7	2	14	✔	✔	✔	
1111 Hours Hindi	2022	108	-	1	1	1	1	5			✔	
Shrutilipi	2022	6k	-	12	2	2	1	1		✔	✔	
WenetSpeech	2022	10k	-	1	4	4	2	10		✔	✔	
Samromur Children	2022	131	3k	1	1	1	1	5		✔		
SDS-200	2022	200	4k	1	3	3	1	2		✔	✔	
aidatatang	2022	200	600	1	1	1	1	7			✔	
Fleurs	2022	1k	-	102	3	3	1	11	✔	✔	✔	
OLKAVS	2022	1k	1k	1	2	2	1	14		✔	✔	
Norwegian Parl.	2022	140	267	1	2	2	1	2			✔	
MagicData-RAMC	2022	180	663	1	4	4	1	15		✔	✔	
Kathbath	2022	2k	1k	12	2	2	1	3		✔	✔	
Hebrew Kan	2022	9	-	1	1	1	1	3				
Hebrew Coursera	2022	36	-	1	1	1	1	7				
Bloom Speech	2022	428	-	56	5	5	1	8	✔	✔		
English-Vietnamese	2022	508	-	2	1	1	1	7			✔	
Earnings-22	2022	119	125	1	1	1	3	2	✔		✔	
YODAS	2023	370k	-	149	3	3	1	1	✔	✔		
AFRISPEECH-200	2023	200	2k	20	14	14	1	6	✔	✔	✔	
Aalto Finnish Parl.	2023	3k	449	1	1	1	1	2		✔		
ReazonSpeech	2023	35k	-	1	2	2	1	1			✔	
EdAcc	2023	40	120	1	1	1	1	8		✔		
RixVox	2023	5k	-	1	1	1	1	2				
Japanese Anime Speech	2023	110	-	1	1	1	1	7				
Snow Mountain	2023	273	11	14	2	2	1	1	✔		✔	

Continued on next page

Table 6: **Audio collections and properties**.

Collection	Year	Property Counts							Category			Perm
		Hr	Spkr	Lang	Creat	Tasks	Src	Top	US	Ac	Ind	Use
Samromur Milljon	2023	967	17k	1	1	1	1	5	✔			🔵
Bud500	2024	500	-	1	1	1	2	4				🔵🔴
VibraVox	2024	18	200	1	1	1	1	1	✔			🔵
M2ASR	Mult.	448	655	4	3	3	1	9	✔			🟡

Table 7: **Video collections and properties**. Collection properties include numbers of hours of video, datasets, creator institutions, countries of creator institutions, and data sources. The Use column indicates whether a collection includes data freely usable even for commercial purposes (🔵), data usable only for noncommercial purposes or academic research (🔴) and data whose license status is not specified precisely enough to allow us to determine commercial use permissions (🟡). Note that each collection may have different datasets with one, two, or all three of these statuses. Finally, the Avail column indicates whether a dataset is available online (✔) or has been taken down, usually for legal reasons (✗). Datasets are sorted chronologically to highlight trends over time.

Collection	Year	Property Counts					Permissions	
		Hours	Datasets	Countries	Creators	Sources	Use	Avail
HOLLYWOOD2	2009	20	1	1	1	1	🟡	✔
Collective	2009	-	1	1	1	1	🟡	✔
HMDB	2011	7k	1	2	3	5	🔵	✔
UCF101	2012	26	1	1	1	1	🟡	✔
YouCook	2013	1k	1	1	1	1	🟡	✔
50 Salads	2013	40	1	1	1	1	🔴	✔
StoryGraphs	2014	7	1	1	1	1	🟡	✔
Hollywood Ext.	2014	9	1	1	1	1	🔵	✔
Breakfast	2014	77	1	2	2	1	🔵	✔
Sports-1M	2014	106k	1	1	1	1	🔵	✔
THUMOS	2014	254	1	2	4	1	🟡	✔
VideoStory	2014	743	1	1	1	1	🟡	✔
SumMe	2014	1	1	2	3	1	🟡	✔
TVSum	2015	4	1	1	1	1	🔵	✔
Volleyball	2015	-	1	1	1	1	🟡	✔
ActivityNet	2015	849	1	2	2	1	🔵	✔
MovieQA	2015	381	1	3	3	1	🟡	✗
Mars	2016	-	1	1	4	1	🟡	✔
NTU RGB+D	2016	74	1	1	1	1	🟡	✔
MSR-VTT	2016	41	1	1	1	1	🟡	✔
Charades	2016	82	1	2	4	1	🟡	✔
VTW	2016	213	1	2	2	1	🟡	✔
Youtube-8M	2016	350k	1	1	1	1	🟡	✔
Narrated Instr. Vid.	2016	7	1	2	4	1	🔵	✔
TGIF	2016	86	1	1	3	1	🟡	✔
MultiTHUMOS	2017	30	1	2	3	1	🔵	✔
ImageNet-Vid	2017	9	1	1	1	1	🔴	✔
PKU-MMD	2017	50	1	1	2	1	🟡	✔
20BN-SOMETHING	2017	121	1	1	1	1	🟡	✔

Continued on next page

Table 7: **Video collections and properties**.

Collection	Year	Property Counts					Permissions	
		Hours	Datasets	Countries	Creators	Sources	Use	Avail
YouCook2	2017	176	1	1	2	1	●	✔
VoxCeleb	2017	2k	1	2	1	1	●	✔
Davis	2017	-	1	1	2	1	●	✔
QFVS	2017	20	1	1	2	1	●	✔
DiDeMo	2018	275	1	1	1	1	●	✔
SOA	2018	2k	1	1	1	1	●	✔
Charades-Ego	2018	69	1	1	1	1	●	✔
EPIC-KITCHENS	2018	100	1	3	3	1	●	✔
MovieGraphs	2018	94	1	1	3	1	●	✗
How2	2018	2k	1	1	1	1	●	✔
VLOG	2018	336	1	1	1	1	●	✔
VaTeX	2019	115	1	2	2	1	●	✔
20BN-jester	2019	13	1	1	1	1	●	✔
HowTo100M	2019	134k	1	2	4	1	●	✔
COIN	2019	476	1	1	2	1	●	✔
MMAct	2019	100	1	2	2	1	●	✔
HACS	2019	833	1	1	3	1	●	✔
CrossTask	2019	376	1	4	5	1	●	✔
Moments in Time	2019	833	1	1	1	11	●	✔
TRECVid	2019	1k	1	1	1	2	●	✔
MSA	2019	516	1	2	2	1	●	✔
Toyota Smarthome	2019	269	1	1	1	1	●	✔
TITAN	2020	3	1	1	1	1	●	✔
VIOLIN	2020	582	1	1	1	1	●	✔
RareAct	2020	21	1	3	5	1	●	✔
TinyVIRAT	2020	11	1	1	1	1	●	✔
100DOH	2020	5k	1	1	2	1	●	✔
Oops!	2020	50	1	1	1	1	●	✔
OmniSource-Web	2020	13k	1	1	1	3	●	✔
Condensed Movies	2020	1k	1	1	1	1	●	✔
MovieScenes	2020	250	1	2	2	1	●	✔
EEV	2020	370	1	1	2	1	●	✔
Movie-Net	2020	3k	1	1	1	1	●	✔
FineGym	2020	708	1	1	1	1	●	✔
HAA500	2020	5	1	2	4	1	●	✔
LEMMA	2020	11	1	1	1	2	●	✔
HVU	2020	96k	1	3	5	1	●	✔
Apes	2021	36	1	3	3	1	●	✔
WebVid	2021	13k	1	2	2	1	●	✗
VideoLT	2021	14k	1	2	4	1	●	✔
HOMAGE	2021	30	1	1	2	1	●	✔
UAV-Human	2021	18	1	2	2	1	●	✔
HD-VILA-100M	2021	372	1	1	1	1	●	✔
M-MiT	2021	833	1	1	1	2	●	✔
Mimetics	2021	1	1	1	1	1	●	✔
Spoken Moments	2021	417	1	1	3	11	●	✔

Continued on next page

Table 7: **Video collections and properties**.

Collection		Property Counts					Permissions	
	Year	Hours	Datasets	Countries	Creators	Sources	Use	Avail
QuerYD	2021	207	1	1	1	2	🟡	✔
MAD	2022	1k	1	1	1	1	🟡	✔
FERV39k	2022	16	1	1	1	1	🔴	✔
CDAD	2022	215	1	1	2	1	🟡	✔
MVBench	2023	-	1	1	6	12	🔵	✔
VidProm	2024	240k	1	2	2	5	🔴	✔
ShareGPT4Video	2024	3k	1	1	4	5	🔴	✔
OpenVid-1M	2024	52k	1	1	3	5	🔵	✔
FineVideo	2024	3k	1	1	1	1	🔵	✔
Disney Vid. Gen.	2024	7	1	1	-	2	🔵	✔
Kinetics	Mult.	4k	3	1	1	2	🟡	✔
Ego4D	Mult.	5k	2	1	2	1	🔵🟡	✔
MPII	Mult.	110	3	1	2	2	🟡	✔
Project-Aria	Mult.	1k	2	1	1	1	🔵	✔
Ava	Mult.	146	2	1	1	2	🔵	✔
LSMDC	Mult.	316	2	4	10	1	🔵🟡	✔