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

Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, Nayan Saxena, Ahmad Mustafa Anis, Emad A. Alghamdi, Vu Minh Chien, 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 (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., 2024c; Malik et al., 2021; Prabhavalkar et al., 2023; Li et al., 2019b), 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 a 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., 2024b; 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).

[1] OpenAI Terms of Use

[2] Llama 3.1 Acceptable Use Policy

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 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., 2024c) 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., 2024a). 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., 2022c); (2) large-scale speech datasets are typically designed for ASR (Li et al., 2023b); (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., 2020b).

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., 2024b; 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

[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.

of non-author video experts to supplement the list with relevant datasets published at CVPR, ICCV, 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

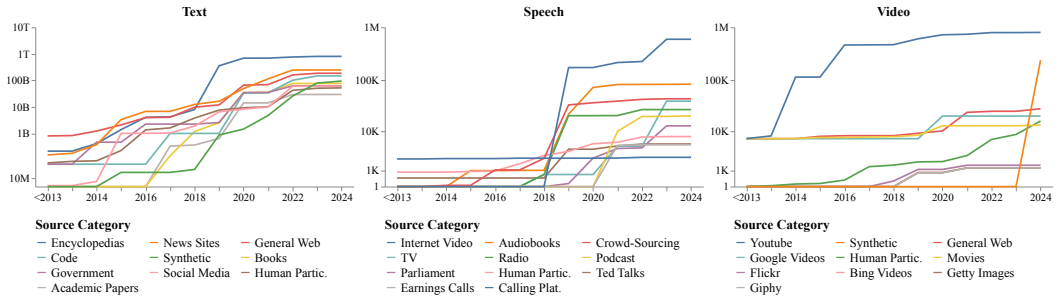


Figure 1: The cumulative size of data (log-scale tokens for text, hours for speech/video) from each source category, across modalities. The source categories in the legend are sorted descending by quantity. **Speech and video sources are increasingly dominated by internet videos and YouTube. Though text is especially web or encyclopedia (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, 2024a) 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).

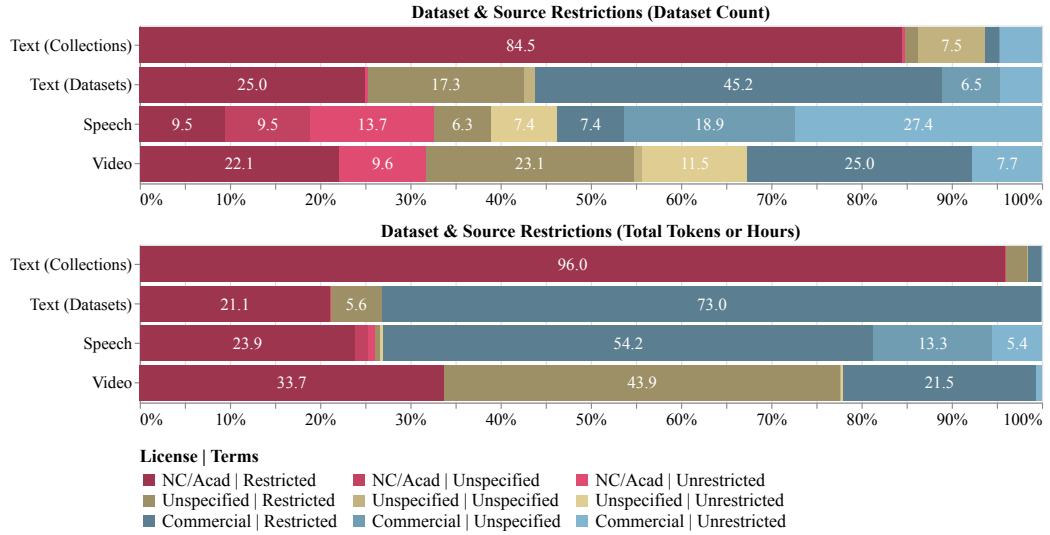


Figure 2: The distribution of restrictions from dataset *licenses* and their sources' *terms*. We break this down by the count of datasets (top), as well as total tokens or hours (bottom). Each license is categorized as Non-commercial/Academic (NC/Acad), Unspecified, or Commercially licensed. Each dataset 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* are a contract between a platform and its users governing 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). How closely practitioners adhere 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.

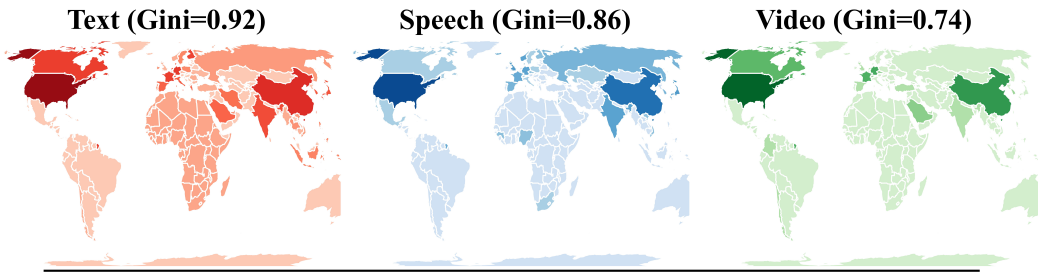
Data source terms are much more restrictive than the dataset's documented license restrictions. In Figure 2, we find only 25%, 33%, and 32% of text/speech/video datasets are licensed non-commercially. This value is even lower if we consider the proportion of tokens or hours, with 21%, 26%, and 33% of 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., 2024a; 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., 2019a; Sigurdsson et al., 2016a). Prior work has measured cultural, ideological, geographic and linguistic imbalances in data (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., 2024a), 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., 2019a)).

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 top-level language families from Glottolog 5.0. 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.[5] 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.

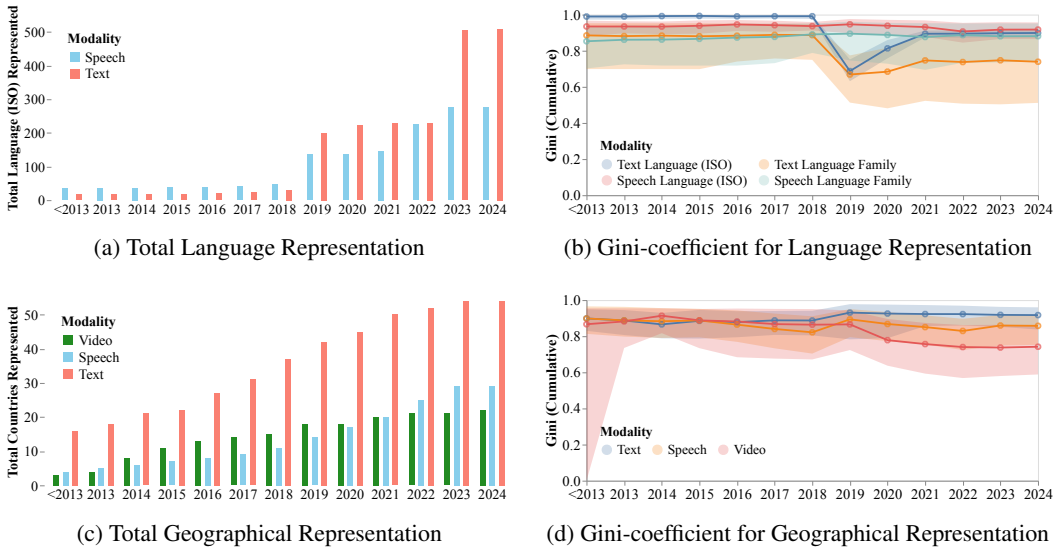
In Figure 3, we also compare the distribution of datasets, and of tokens or hours by continent. Dataset creators affiliated with African or South American organizations account for fewer than 0.2% of all tokens or hours, in each modality. In contrast, Asian affiliated organizations represent large proportions of the data, particularly for speech (39% of hours, attributed predominantly to YODAS (Li et al., 2023b)). 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., 2024a). 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. Subsequently, newer datasets remain predominantly monolingual, causing measures of concentration in text languages, speech languages, and language families to remain consistently high.

Academia, research non-profits, and industry labs continue to drive public dataset development. We manually categorize the organizations creating popular datasets into: Academic Organization

[5]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.**

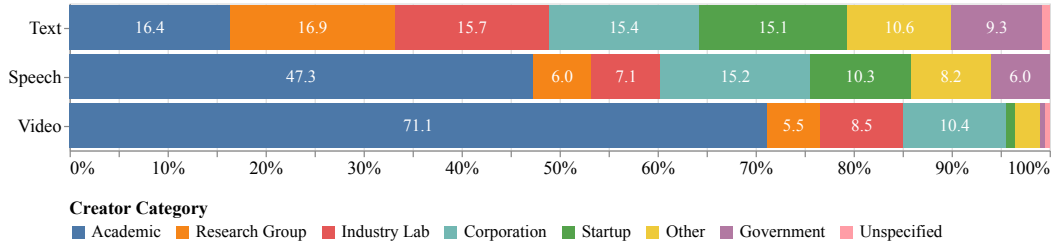


Figure 5: The distribution of creator organizations by modality. **Most public speech and video datasets are developed by academic organizations, whereas text datasets are developed by a wide mix of academia, non-profit or industry labs, as well as startups.**

(e.g., universities), Research Groups (e.g., non-profits like BigScience, EleutherAI and AI2), Industry Labs (e.g., Cohere for AI, Google DeepMind), Corporations (e.g. Google, Meta), Startups (e.g., OpenAI, Anthropic), Governments, Unspecified (owner affiliation not shared), and 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. Academic organizations' greater role in Video and Speech may suggest special impediments (e.g., privacy-related) to commercial releases of these datasets.

4 Discussion

The rise of web, social media, and synthetic data may pose greater privacy, copyright, and bias risks. Section 3.1 discusses the rise of web content and particularly social media as primary sources for speech and video. Figure 1 shows these sources now exceed by at least an order of magnitude more traditional, curated sources like movies, audiobooks, radio, TV, or data produced by human participants. Their mostly user-generated content makes these websites a natural source for the quantity, freshness, and heterogeneity needed to train general-purpose models (Longpre et al.,

2023; Aghajanyan et al., 2023). However, prior work suggests that user-generated web data is also more challenging to use than curated data, particularly for privacy, copyright, bias, 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., 2023c), creating privacy and copyright concerns (Zhang et al., 2023a). 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 sites are a very prominent data source, but their terms often restrict crawling or commercial use. We find 71% of Video and 69% of Speech data to be from YouTube, whose scale, freshness, and multimodality (containing videos, speech, images, and text) have made it a prominent data source (Abu-El-Haija et al., 2016b; Aytar et al., 2018; Chang et al., 2020; Uthus et al., 2023; Coats, 2023; Li et al., 2023b). However, YouTube is a social media platform owned by Google and its Terms of Service[6] prohibit third-party crawling. While content creators retain ownership of material uploaded to YouTube, the YouTube terms of service also grant Google a license to reproduce, modify, display, and use the content for YouTube's "business" (which may include building machine learning models), and forbid crawling by third parties, even if the copyright holder has selected a permissive license. Model developers like 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 sites like Reddit and StackOverflow[7] have also recently adopted restrictive terms, with access being restricted just as these data sources become critical to scaling AI systems. The enforceability of these licenses and terms, however, is an open legal question beyond the scope of our work.

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 reach 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 separating and identifying appropriate datasets remains difficult across these modalities. Without continued audits and documentation, practitioners may have to either take on avoidable risk or forgo large collections of partially viable data, hampering data scaling laws (Kaplan et al., 2020). We hope our audit will help practitioners apply their own standards and make informed decisions on training data use.

The limitations of measures of geographical and linguistic representation. Measures of geographic and linguistic representation are imperfect: We have only partial information about the developers' identities (including for privacy reasons), limited transparency into how frequently these datasets are used, and cannot say how far 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 *public* datasets. Indeed, we contend that it is necessary to measure representation in AI data to understand progress, or its absence, towards AI systems that fairly serve the broad user community. Figures 3 and 4 show that despite the absolute rise of geographic and linguistic representation, a western-centric skew persists across thousands of surveyed datasets. We release all audit materials for transparency, replicability, and further research use.

Conducting representative analyses of an ecosystem comes with assumptions. The AI ecosystem is decentralized. Text datasets, for example, are often hosted on HuggingFace, unlike Speech and Video. Similarly, while Text data is frequently repackaged for general-purpose post-training, this is less true of other modalities. Scoping and dataset selection thus need to be modality-specific, rather than a single general protocol unable to capture these nuances. Similarly, we studied modalities of interest to foundation model development (Brooks et al., 2024; Radford et al., 2023), but many others remain for future work (e.g., images, 3D shapes, tabular, time series, graphs, and geospatial data).

[6] YouTube Terms of Service.

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

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.

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., 2020a; 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.

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.

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.

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.

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.

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 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.

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.

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

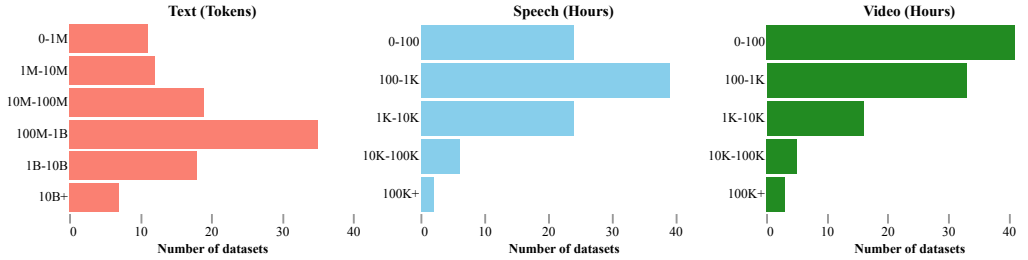


Figure 6: The distribution of dataset sizes for each modality. Most text data collections are between 100M-1B tokens. **Speech datasets average 100-1k hours, and video datasets are usually the smallest, commonly less than 100 hours.**

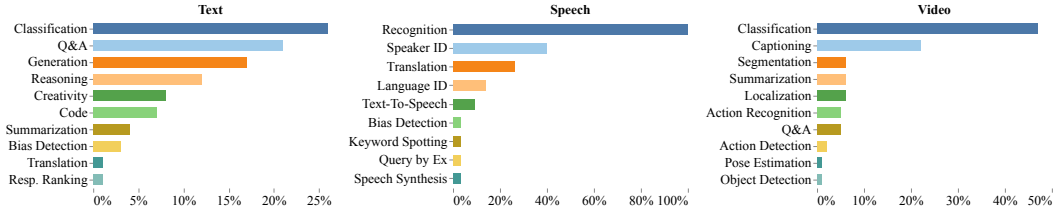


Figure 7: The task distribution of datasets, across modalities. Post-training text and video datasets are predominantly based on classification. For text, generation and reasoning are rising categories. All speech datasets are recognition-based, particularly for speaker, language, or in the process of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

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 in Appendix F.

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 (Rajpurkar et al., 2016a) 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. (2019a; 2021a). 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	🤖	🔵🔴	✔
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	🤖	🔵	

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
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	✔		✔	🔴
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	✔	✔		🔵

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
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		✔		🟡
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		✔	✔	🟡

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	
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	✔		✔	●	
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		PROPERTY COUNTS					PERMISSIONS	
	YEAR	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	🟡	✔
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	🔴	✔

Continued on next page

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

COLLECTION		PROPERTY COUNTS					PERMISSIONS	
	YEAR	HOURS	DATASETS	COUNTRIES	CREATORS	SOURCES	USE	AVAIL
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	●	✔
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	●	✔

Continued on next page

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

COLLECTION		PROPERTY COUNTS					PERMISSIONS	
	YEAR	HOURS	DATASETS	COUNTRIES	CREATORS	SOURCES	USE	AVAIL
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		✔

E CONTRIBUTIONS

Here we break down contributions to this work. Contributors are listed alphabetically, except for team leads who are placed first.

- **Text Datasets** Shayne Longpre (lead), Jad Kabbara (lead), Ahmad Anis, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Kun Qian, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Nayan Saxena, Niklas Muennighoff, Naana Obeng-Marnu, Robert Mahari, Seonghyeon Ye, Seungone Kim, Shayne Longpre, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, William Brannon, Xuhui Zhou, Yizhi Li, An Dinh, Caroline Chitongo, Christopher Klamm, Da Yin, Damien Sileo, Ariel Lee
- **Reviewing Text Dataset Metadata** Jad Kabbara (lead), Shayne Longpre (lead), Robert Mahari, Damien Sileo, Niklas Muennighoff, William Brannon,
- **Data Explorer Features** Shayne Longpre (lead), Christopher Klamm, Vu Minh Chien,
- **Speech Datasets** Nikhil Singh (lead), Manuel Cherep (lead), An Dinh, Minnie Liang, Shrestha Mohanty
- **Video Datasets** Kush Tiwary (lead), Joanna Materzynska (lead), Vivek Sharma, Shayne Longpre, Robert Mahari, Jad Kabbara, William Brannon, Tobin South, Shrestha Mohanty, Nikhil Singh, Manuel Cherep
- **Data Analysis** Shayne Longpre (lead), Nikhil Singh (lead), Manuel Cherep (lead), Kush Tiwary (lead), Joanna Materzynska (lead), Naana Obeng-Marnu (lead), William Brannon (lead),
- **Writing** Shayne Longpre (lead), Jad Kabbara (lead), Nikhil Singh, Manuel Cherep, Kush Tiwary, Joanna Materzynska, Robert Mahari
- **Legal Analysis** Robert Mahari (lead), Luis Villa
- **Visualizations & Visual Data Analysis** Nikhil Singh (lead), Manuel Cherep (lead), Kush Tiwary (lead), Joanna Materzynska (lead), Naana Obeng-Marnu (lead), William Brannon (lead), Shayne Longpre (lead), Ariel Lee, Hamidah Oderinwale, Campbell Lund
- **Senior Advisors** Stella Biderman, Sara Hooker, Jad Kabbara, Sandy Pentland, Luis Villa, Caiming Xiong

F ATTRIBUTION CARD

Here we provide detailed information about the licenses of each data collection and its constituent datasets, and cite all of the papers (455 in all) which introduced datasets we consider. Text datasets are laid out in Table 8, audio datasets in Table 9, and video datasets in Table 10. Because of the large number of references, we include a second bibliography after the tables (named 'Attribution Card References'), with numbered citations in this section referring to that second bibliography.

Table 8: **References and licenses for alignment-tuning (text)** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
10k Prompt Ranked	Unspecified	–
AgentInstruct	Unspecified, CC BY 4.0, MIT License	Shridhar et al. (2021); Yao et al. (2023); Liu et al. (2023c); Zeng et al. (2023); Deng et al. (2023)
Aya	Apache License 2.0	Singh et al. (2024b)
Bactrian-X	CC BY-SA 3.0, CC BY-NC 4.0	Li et al. (2023a)
COBRA Frames	BigScience OpenRAIL-M	Zhou et al. (2023b)
COIG	Various	Zhang et al. (2023b); Bai et al. (2024a)
Capybara	Various	–
ChatDoctor	Unspecified	Li et al. (2023d)
ChatbotArena	CC BY 4.0, CC BY-NC 4.0	Zheng et al. (2023)
Cidar	CC BY-NC 4.0	Alyafeai et al. (2024)
CollectiveCognition	MIT License	–
Conifer	Apache License 2.0	Sun et al. (2024)
Deita 10K	Apache License 2.0, CC BY-NC 4.0	Liu et al. (2024b)

Continued on next page

Table 8: **References and licenses for alignment-tuning (text)** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
DialogStudio	Various	Chen et al. (2021a); Wei et al. (2018); Lin et al. (2021b); Chawla et al. (2021); He et al. (2018); Mrksic et al. (2017); Qian et al. (2022); Liu et al. (2021); El Asri et al. (2017); Quan et al. (2019); Chen et al. (2019; 2022b); Eric & Manning (2017); Zang et al. (2020); Shalyminov et al. (2019); Martin et al. (2020); Peskov et al. (2019); Eric et al. (2019); Moon et al. (2019); Rastogi et al. (2020); Mosig et al. (2020); Chiu et al. (2022); Shah et al. (2018); Byrne et al. (2019); Mrkšić & Vulić (2018); Shang et al. (2018); Rameshkumar & Bailey (2020); Fabbri et al. (2021); Chen et al. (2021c); Mukherjee et al. (2022); Shang et al. (2018); Zhu et al. (2021); Zhong et al. (2021); Gliwa et al. (2019); Chen et al. (2022a); Feigenblat et al. (2021); Li et al. (2019c); Dinan et al. (2019a); Rashkin et al. (2019); Bai et al. (2022); Chen et al. (2023a); Kim et al. (2022); Myers et al. (2020); Reddy et al. (2019); Yu et al. (2019a); Talmor & Berant (2018); Nan et al. (2021; 2022); Gu et al. (2021); Chen et al. (2020b); Gupta et al. (2018a); Li et al. (2021a); Talmor et al. (2021); Yu et al. (2019c); Iyyer et al. (2017); Yu et al. (2019b); Parikh et al. (2020); Yih et al. (2016); Zhong et al. (2017); Pasupat & Liang (2015); Komeili et al. (2022); Dinan et al. (2019b); Hemphill et al. (1990); Casanueva et al. (2020); Zhang et al. (2022b); Larson et al. (2019); Rastogi et al. (2020); Liu et al. (2019; 2013); Coope et al. (2020); Coucke et al. (2018); Gupta et al. (2018b)

Continued on next page

Table 8: **References and licenses for alignment-tuning (text)** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
Dynosaur	Various	Adlakha et al. (2022); Agarwal et al. (2021); Akyürek et al. (2022); Amini et al. (2019); Ardanuy et al. (2020); Austin et al. (2021); Azerbayev et al. (2023); Bai et al. (2022); Bajaj et al. (2018); Balakrishnan et al. (2019); Bartolo et al. (2020); Bisk et al. (2019); Boratko et al. (2020); Botha et al. (2018); Boudin & Gallina (2021); Bravo et al. (2015); Brown et al. (2020b); Byrne et al. (2019); Cao & Wang (2021); Cao et al. (2022); Casanueva et al. (2020); Cetoli et al. (2019); Chalkidis et al. (2019b;a; 2021); Chan et al. (2022); Chapuis et al. (2021); Chen et al. (2020a); Cheng et al. (2022); Chouldechova (2017); Christmann et al. (2019); Clark et al. (2019; 2018); Cobbe et al. (2021); Coucke et al. (2018); Dankers et al. (2022); Dasigi et al. (2019); Devaraj et al. (2021); DeYoung et al. (2021); Diggelmann et al. (2021); Emelin et al. (2020); Fabbri et al. (2019); Faruqui & Das (2018); Feng et al. (2021); Gallina et al. (2019); Ganesan et al. (2010); Gazzola et al. (2019); George & Mamidi (2019); Geva et al. (2019); Gliwa et al. (2019); Gorrell et al. (2018); Gu et al. (2022); Gupta et al. (2021); Ha & Eck (2017); Haagsma et al. (2020); Hazoom et al. (2021); Henderson et al. (2022); Hendrycks et al. (2021); Huang et al. (2019; 2021); Huang (2022); Irwin et al. (2020); Ivgi et al. (2022); Iyer et al. (2017); Jiang et al. (2020; 2021); Jin et al. (2019); Joshi et al. (2017); Juraska et al. (2019); Jurczyk et al. (2016); Kanade et al. (2020); Kaushik et al. (2020); Khot et al. (2020; 2018); Kim et al. (2018); Kornilova & Eidelman (2019); Kury et al. (2020); Lai et al. (2017); Lake & Baroni (2018); Lebret et al. (2016); Lewis et al. (2017); Li et al. (2022; 2019a); Lin et al. (2020a)

Continued on next page

Table 8: **References and licenses for alignment-tuning (text)** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
Dynosaur (cont'd)	Various	Lin et al. (2020b; 2019; 2022); Ling et al. (2017); Liu et al. (2019); Louis et al. (2020); Lowe et al. (2016); Malo et al. (2013); Martin et al. (2018); Merity et al. (2016); Mihaylov et al. (2018); Mishra et al. (2023); Moniz & Torgo (2018); Mostafazadeh et al. (2020); Nan et al. (2021); Narayan et al. (2018); Nguyen et al. (2021); Nie et al. (2020); Novikova et al. (2017); Paik et al. (2021); Pakhomov et al. (2010); Pang & Lee (2005); Pavlichenko et al. (2021); Pedersen et al. (2007); Perez-Beltrachini et al. (2019); Petroni et al. (2019); Pham et al. (2023); Rajani et al. (2019); Rajpurkar et al. (2016b); Rameshkumar & Bailey (2020); Rashkin et al. (2019); Rastogi et al. (2020); Royer et al. (2018); Rush et al. (2015); Rust et al. (2023); Saeidi et al. (2018); Saha et al. (2018); Sakaguchi et al. (2019); Sanh et al. (2022); Sap et al. (2019); Schulz et al. (2020); See et al. (2017); Sharma et al. (2019); Shriberg et al. (1998); Sileo & Moens (2023); Soleimani et al. (2021); Stolcke et al. (2000); Tafjord et al. (2019; 2018); Talmor et al. (2019); Tandon et al. (2019); Tang et al. (2020); Thawani et al. (2021); Thorne et al. (2018); Tyleček & Šára (2013); Ullrich et al. (2023); Ushio et al. (2023); Wang et al. (2022a; 2020c; 2019; 2023a); Warstadt et al. (2023); Welbl et al. (2018; 2017); Weller et al. (2020); Weston et al. (2015); Williams et al. (2020); Wu et al. (2018); Xiong et al. (2019a); Yang et al. (2018); Yu et al. (2019b); Zellers et al. (2019); Zhang et al. (2016; 2023c; 2019); Zhou et al. (2019; 2023a); Zhu et al. (2022)
EverythingLM	MIT License	–
ExpertQA	MIT License	Malaviya et al. (2024)
Feedback Coll.	MIT License	Kim et al. (2024)

Continued on next page

Table 8: **References and licenses for alignment-tuning (text)** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
Glaive Code Asst.	Apache License 2.0	–
Gretel Text-to-SQL	Apache License 2.0	–
HelpSteer	CC BY 4.0	Wang et al. (2023b)
Indic-Instr.	Various	Gala et al. (2024)
InstAr	Various	Hu et al. (2020); Einea et al. (2019); Mozannar et al. (2019); Pratapa et al. (2022); Chouikhi et al. (2024); Abbas et al. (2011); Abdelghany et al. (2020); Orabi et al. (2020); ElSahar & El-Beltagy (2015); Elnagar & Einea (2016); Pieri et al. (2024); Alghamdi et al. (2022); Abdallah et al. (2024); Biltawi et al. (2020); Aloui et al. (2024); El-khair (2016)
KIWI	CC BY-SA 4.0	Xu et al. (2024a)
LMSYS-Chat-1M	LMSYS-Chat-1M Dataset License, Anthropic, Llama 2	Zheng et al. (2024a)
Llama2-MedTuned-Instr.	CC BY-NC 4.0	Rohanian et al. (2023)
LongAlign-10k	Anthropic, Apache License 2.0	Bai et al. (2024b)
Magpie-Pro	Meta Llama3 Community License	Xu et al. (2024b)
MathDial	CC BY-SA 4.0, MIT License	Macina et al. (2023)
MathInstr.	MIT License	Yue et al. (2023)
MedInstr.	Unspecified	Zhang et al. (2024)
Medical Meadow	Various	Han et al. (2023); Wang et al. (2020b); Jin et al. (2020); Savery et al. (2020)
MegaWika	CC BY-SA 4.0	Barham et al. (2023)
MetaMathQA	MIT License	Yu et al. (2023)
Nectar	Various	–
No Robots	CC BY-NC 4.0	–
Open Asst. v2	Apache License 2.0	Köpf et al. (2023)
Open-Platypus	Various	Sawada et al. (2023); Dettmers et al. (2023); Lightman et al. (2023); Yu et al. (2020); Wang et al. (2024); Lu et al. (2022); Chen et al. (2023b)
OpenGPT Healthcare	Unspecified, OGL 3.0	–
OpenMathInstr.-1	Custom, MIT License, Apache License 2.0	Toshniwal et al. (2024)
Orca-Math	Various	Mitra et al. (2024)
PII-Masking-200k	Non Commercial	–
PMC-LLaMA Instr.	Unspecified, Apache License 2.0	Wu et al. (2023); Jin et al. (2019)
Pure-Dove	Apache License 2.0	–
PygmalionAI-PIPPA	Apache License 2.0	Gosling et al. (2023)

Continued on next page

Table 8: **References and licenses for alignment-tuning (text)** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
Reasoning	Apache License 2.0	–
RiddleSense	MIT License	Lin et al. (2021a)
SeaBench	Apache License 2.0	Nguyen et al. (2023)
SelFee	MIT License	Ye et al. (2023)
Synth.-GSM8K-Refl.	Meta Llama3 Community License	–
Thai Gen AI	Various	–
UltraChat-200k	CC BY-NC 4.0	Ding et al. (2023)
UltraFeedback Argilla	Various	–
WildChat	AI2 ImpACT License - Low Risk	Zhao et al. (2023)

Table 9: **References and licenses for audio** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
1111 Hours Hindi	Custom	Bhanushali et al. (2022)
120h Spanish Speech	CC0 1.0	–
AFRISPEECH-200	CC BY-NC-SA 4.0	Olatunji et al. (2023)
AISHELL-1	Apache 2.0	Bu et al. (2017)
AISHELL-2	Unspecified	Du et al. (2018)
AISHELL-4	CC BY-SA 4.0	Fu et al. (2021)
ALLSSTAR	CC BY 4.0	Bradlow (2010)
AMI	CC BY 4.0	Carletta et al. (2006)
Aalto Finnish Parl.	Custom	Virkkunen et al. (2022)
African Acc. French	Apache 2.0	–
Basq., Cat. and Gal.	CC BY-SA 4.0	Kjartansson et al. (2020)
Bloom Speech	Various	Leong et al. (2022)
Bud500	Apache 2.0, CC BY-NC-SA 4.0	–
CSJ	Custom	Maekawa (2003)
CSLU 1.2	CSLU Agreement	Lander, T (2007)
CSLU 22 Langs.	CSLU Agreement	Lander, T (2005)
ClarinPL	CC BY 4.0	Korzinek et al. (2017)
CoNASE	Custom	Coats (2019)
CoVoST-2	CC0 1.0	Wang et al. (2020a)
Common Voice 17	CC0 1.0	Ardila et al. (2020a)
Crowd Sourced Speech	CC BY-SA 4.0	Kjartansson et al. (2018)
Czech Parliament	CC BY 4.0	Kratochvil et al. (2020)
DiDiSpeech	Unspecified	Guo et al. (2021)
Earnings-22	Unspecified	Del Rio et al. (2022)
EdAcc	CC BY-SA 4.0	Sanabria et al. (2023)
Eng. Acc. in Brit. Isles	CC BY-SA 4.0	Demirsahin et al. (2020)
English-Vietnamese	CC BY-NC-ND 4.0	Nguyen et al. (2022)
FT SPEECH	Custom	Kirkedal et al. (2020)
Fisher	LDC User Agreement	Cieri et al. (2004)
Fleurs	CC BY 4.0	Conneau et al. (2022)
GigaSpeech	Apache 2.0	Chen et al. (2021b)
Golos	Custom	Karpov et al. (2021)
Hebrew Coursera	Unspecified	–
Hebrew Kan	Unspecified	–
Highland Puebla Nahuatl	CC BY-NC-SA 3.0	Shi et al. (2021)
JTUBESPEECH	Unspecified	Takamichi et al. (2021)
Japanese Anime Speech	CC0 1.0	–
KSC	CC BY 4.0	Khassanov et al. (2021)
Kathbath	CC0 1.0	Javed et al. (2022)
KeSpeech	Custom	Tang et al. (2021)
KsponSpeech	Unspecified	Bang et al. (2020)

Continued on next page

Table 9: **References and licenses for audio** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
LJSpeech	Public Domain	Ito & Johnson (2017)
LaboroTVSpeech	Custom	Ando & Fujihara (2021)
LibriSpeech	CC BY 4.0	Panayotov et al. (2015)
M-AILABS	Custom	Solak (2024)
M2ASR	Unspecified	Shi et al. (2017); Mamtimin et al. (2023); Zhi et al. (2021); Li et al. (2017)
MAGICDATA	CC BY-NC-ND 4.0	–
MASC	CC BY 4.0	Al-Fetyani et al. (2023)
MaSS	Unspecified	Boito et al. (2020)
MagicData-RAMC	CC BY-NC-ND 4.0	Yang et al. (2022)
MediaSpeech	CC BY 4.0	Kolobov et al. (2021)
Minds14	CC BY 4.0	Gerz et al. (2021)
MuST-C	CC BY-NC-ND 4.0	Di Gangi et al. (2019)
Multiling. LibriSpeech	CC BY 4.0	Pratap et al. (2020)
Multiling. TEDx	CC BY-NC-ND 4.0	Salesky et al. (2021)
NST Danish	CC0 1.0	–
NST Norwegian	CC0 1.0	–
NST Swedish	CC0 1.0	–
Nigerian English	CC BY-SA 4.0	–
Norwegian Parl.	CC0 1.0	Solberg & Ortiz (2022)
Norwegian Parl. Speech	CC0 1.0	Solberg & Ortiz (2022)
OLKAVS	Custom	Park et al. (2023)
OpenSTT	CC BY-NC 4.0	Andrusenko et al. (2020)
People's Speech	Various	Galvez et al. (2021)
QASR	Unspecified	Mubarak et al. (2021)
RTVE	Custom	–
ReazonSpeech	CDLA-Sharing-1.0	Yin et al. (2023)
Regional Af. Am. Lang.	CC BY-NC-SA 4.0	–
RixVox	CC BY 4.0	–
SDS-200	Custom	Plüss et al. (2022)
SPGISpeech	Custom	O'Neill et al. (2021)
Samromur	CC BY 4.0	Mollberg et al. (2020)
Samromur Children	CC BY 4.0	Hernandez Mena et al. (2022)
Samromur Milljon	CC BY 4.0	–
Shrutilipi	CC0 1.0	Bhogale et al. (2022)
Snow Mountain	CC BY-SA 4.0	Raju et al. (2023)
Spoken Wikipedia	CC BY-SA 4.0	Baumann et al. (2019)
Switchboard	LDC User Agreement	Godfrey et al. (1992)
TED-LIUM3	CC BY-NC-ND 3.0	Hernandez et al. (2018)
THCHS-30	Apache 2.0	Wang & Zhang (2015)
THUYG-20	Apache 2.0	Rozi et al. (2015)

Continued on next page

Table 9: **References and licenses for audio** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
TIMIT	LDC User Agreement	Garofolo, John S. et al. (1993)
VCTK	CC BY 4.0	–
VibraVox	CC BY 4.0	–
VoxPopuli	CC0 1.0	Wang et al. (2021)
Vystadial	CC BY-SA 3.0	Korvas et al. (2014)
WenetSpeech	CC BY 4.0	Zhang et al. (2022a)
West Afr. Radio	CC BY-SA 4.0	Doumbouya et al. (2021)
West Afr. Virt. Asst.	CC BY-SA 4.0	Doumbouya et al. (2021)
YODAS	CC BY 3.0	Li et al. (2023c)
Zeroth-Korean	CC BY 4.0	–
aidatatang	CC BY-NC-ND 4.0	–

Table 10: **References and licenses for video** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
100DOH	Custom	Shan et al. (2020)
20BN-SOMETHING	Custom	Goyal et al. (2017)
20BN-jester	Custom	Materzynska et al. (2019)
50 Salads	CC BY-NC-SA 4.0	Stein & McKenna (2013)
ActivityNet	MIT License	Heilbron et al. (2015)
Apes	Unspecified	Alcazar et al. (2021)
Ava	CC BY 4.0	Roth et al. (2019); Gu et al. (2018)
Breakfast	CC BY 4.0	Kuehne et al. (2014)
CDAD	Unspecified	Xiang et al. (2022)
COIN	Custom	Tang et al. (2019)
Charades	Custom	Sigurdsson et al. (2016b)
Charades-Ego	Custom	Sigurdsson et al. (2018)
Collective	Unspecified	Wongun Choi et al. (2009)
Condensed Movies	CC BY 4.0	Bain et al. (2020)
CrossTask	Unspecified	Zhukov et al. (2019)
Davis	Custom	Perazzi et al. (2016)
DiDeMo	BSD 2-Clause License	Hendricks et al. (2018)
Disney Vid. Gen.	Apache 2.0	–
EEV	CC BY 4.0	Sun et al. (2021)
EPIC-KITCHENS	CC BY-NC 4.0	Damen et al. (2018)
Ego4D	Custom, MIT License	Grauman et al. (2022)
FERV39k	CC BY-NC 4.0	Wang et al. (2022b)
FineGym	CC BY-NC 4.0	Shao et al. (2020)
FineVideo	CC BY 4.0	–
HAA500	Unspecified	Chung et al. (2021)
HACS	Custom	Zhao et al. (2019)
HD-VILA-100M	Custom	Xue et al. (2022)
HMDB	CC BY 4.0	Kuehne et al. (2011)
HOLLYWOOD2	Unspecified	Marszalek et al. (2009)
HOMAGE	Unspecified	Rai et al. (2021)
HVU	Custom	Diba et al. (2020)
Hollywood Ext.	MIT License	Bojanowski et al. (2014)
How2	Various	Sanabria et al. (2018)
HowTo100M	Unspecified	Miech et al. (2019)
ImageNet-Vid	CC BY-NC 4.0	Russakovsky et al. (2015)
Kinetics	Unspecified	Kay et al. (2017); Carreira et al. (2018); Smaira et al. (2020)
LEMMA	Unspecified	Jia et al. (2020)
LSMDC	Custom, MIT License	Rohrbach et al. (2016a); Sharma et al. (2020)
M-MiT	Unspecified	Monfort et al. (2021b)

Continued on next page

Table 10: **References and licenses for video** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
MAD	Custom	Soldan et al. (2022)
MMAct	Custom	Kong et al. (2019)
MPII	Unspecified, Custom	Rohrbach et al. (2016b; 2015)
MSA	Unspecified	Xiong et al. (2019b)
MSR-VTT	Unspecified	Xu et al. (2016)
MVBench	MIT License	Li et al. (2024)
Mars	Unspecified	Zheng et al. (2016)
Mimetics	Unspecified	Weinzaepfel & Rogez (2021)
Moments in Time	Custom	Monfort et al. (2019b)
Movie-Net	Unspecified	Huang et al. (2020)
MovieGraphs	Custom	Vicol et al. (2018)
MovieQA	Unspecified	Tapaswi et al. (2016)
MovieScenes	Unspecified	Rao et al. (2020)
MultiTHUMOS	CC BY 4.0	Yeung et al. (2017)
NTU RGB+D	Custom	Shahroudy et al. (2016)
Narrated Instr. Vid.	MIT License	Alayrac et al. (2016)
OmniSource-Web	Apache License 2.0	Duan et al. (2020)
Oops!	CC BY-NC-SA 4.0	Epstein et al. (2020)
OpenVid-1M	CC-BY-4.0	Nan et al. (2024)
PKU-MMD	Unspecified	Liu et al. (2017)
Project-Aria	Apache License 2.0	Pan et al. (2023); Lv et al. (2024)
QFVS	Unspecified	Sharghi et al. (2017)
QuerYD	Unspecified	Oncescu et al. (2021)
RareAct	Unspecified	Miech et al. (2020)
SOA	Unspecified	Diba et al. (2020)
ShareGPT4Video	Attribution-NonCommercial 4.0 International	Chen et al. (2024)
Spoken Moments	Custom	Monfort et al. (2021a)
Sports-1M	CC BY 3.0	Karpathy et al. (2014)
StoryGraphs	Unspecified	Tapaswi et al. (2014)
SumMe	Unspecified	Gygli et al. (2014)
TGIF	Custom	Li et al. (2016)
THUMOS	Custom	Idrees et al. (2017)
TITAN	Non Commercial	Malla et al. (2020)
TRECVid	CC BY-NC-SA 4.0	Awad et al. (2020)
TVSum	CC BY 3.0	Yale Song et al. (2015)
TinyVIRAT	Unspecified	Demir et al. (2020)
Toyota Smarthome	Custom	Das et al. (2019)
UAV-Human	Custom	Li et al. (2021b)
UCF101	Unspecified	Soomro et al. (2012)
VIOLIN	Unspecified	Liu et al. (2020)
VLOG	Custom	Fouhey et al. (2017)

Continued on next page

Table 10: **References and licenses for video** dataset collections presented in this paper. Collections containing material under more than three distinct licenses are marked as having "Various" licenses, and we refer readers to our raw data for the full details. Datasets are sorted alphabetically for ease of dataset lookup.

Collection	Licenses	Cite
VTW	Unspecified	Zeng et al. (2016)
VaTeX	CC BY 4.0	Wang et al. (2020d)
VidProm	CC-BY-NC 4.0	Wang & Yang (2024b)
VideoLT	Non Commercial	Zhang et al. (2021)
VideoStory	Unspecified	Habibian et al. (2014)
Volleyball	Unspecified	Ibrahim et al. (2016)
VoxCeleb	Custom	Nagrani et al. (2018)
WebVid	Custom	Bain et al. (2022)
YouCook	Unspecified	Das et al. (2013)
YouCook2	MIT License	Zhou et al. (2017)
Youtube-8M	Unspecified	Abu-El-Haija et al. (2016a)

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.

Mourad Abbas, Kamel Smaïli, and D. Berkani. Evaluation of Topic Identification Methods on Arabic Corpora. *Journal of Digital Information Management*, 9(5), October 2011. URL https://www.dline.info/fpaper/jdim/v9i5/1.pdf.

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. ArabicaQA: A Comprehensive Dataset for Arabic Question Answering, March 2024. URL http://arxiv.org/abs/2403.17848. arXiv:2403.17848 [cs].

Ahmed Abdelghany, Hammam Abdelaal, Abdulrahman Kamr, and Passent Elkafrawy. Doc2Vec: An approach to identify Hadith Similarities. *Australian Journal of Basic and Applied Sciences*, pp. 46–53, December 2020. doi: 10.22587/ajbas.2020.14.12.5.

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, September 2016a. URL http://arxiv.org/abs/1609.08675. arXiv:1609.08675 [cs].

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*, 2016b.

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.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching, February 2022. URL http://arxiv.org/abs/2110.00768. arXiv:2110.00768 [cs].

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training, March 2021. URL http://arxiv.org/abs/2010.12688. arXiv:2010.12688 [cs].

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.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. Towards Tracing Factual Knowledge in Language Models Back to the Training Data, October 2022. URL http://arxiv.org/abs/2205.11482. arXiv:2205.11482 [cs].

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. MASC: Massive Arabic Speech Corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1006–1013, Doha, Qatar, January 2023. IEEE. ISBN 979-8-3503-9690-4. doi: 10/gtsqzj. URL https://ieeexplore.ieee.org/document/10022652/.

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised Learning from Narrated Instruction Videos, June 2016. URL http://arxiv.org/abs/1506.09215. arXiv:1506.09215 [cs].

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.

Juan Leon Alcazar, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbelaez, Bernard Ghanem, and Fabian Caba Heilbron. APES: Audiovisual Person Search in Untrimmed Video, June 2021. URL http://arxiv.org/abs/2106.01667. arXiv:2106.01667 [cs].

Reem Alghamdi, Zhenwen Liang, and Xiangliang Zhang. ArMATH: a Dataset for Solving Arabic Math Word Problems. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 351–362, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.37.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 101 Billion Arabic Words Dataset, April 2024. URL http://arxiv.org/abs/2405.01590. arXiv:2405.01590 [cs].

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.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-Shaibani. CIDAR: Culturally Relevant Instruction Dataset For Arabic, February 2024. URL http://arxiv.org/abs/2402.03177. arXiv:2402.03177 [cs].

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms, May 2019. URL http://arxiv.org/abs/1905.13319. arXiv:1905.13319 [cs].

Shintaro Ando and Hiromasa Fujihara. Construction of a Large-scale Japanese ASR Corpus on TV Recordings, March 2021. URL http://arxiv.org/abs/2103.14736. arXiv:2103.14736 [cs, eess].

Andrei Andrusenko, Aleksandr Laptev, and Ivan Medennikov. Exploration of End-to-End ASR for OpenSTT – Russian Open Speech-to-Text Dataset. In *Lecture Notes in Computer Science*, volume 12335, pp. 35–44. Springer, Cham, 2020. doi: 10.1007/978-3-030-60276-5_4. URL http://arxiv.org/abs/2006.08274. arXiv:2006.08274 [cs, eess].

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. Living Machines: A study of atypical animacy, November 2020. URL http://arxiv.org/abs/2005.11140. arXiv:2005.11140 [cs].

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 Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus, March 2020a. URL http://arxiv.org/abs/1912.06670. arXiv:1912.06670 [cs].

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 2020b. 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.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models, August 2021. URL http://arxiv.org/abs/2108.07732. arXiv:2108.07732 [cs].

George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quenot. TRECVID 2019: An Evaluation Campaign to Benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & Retrieval, September 2020. URL http://arxiv.org/abs/2009.09984. arXiv:2009.09984 [cs].

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.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. ProofNet: Autoformalizing and Formally Proving Undergraduate-Level Mathematics, February 2023. URL http://arxiv.org/abs/2302.12433. arXiv:2302.12433 [cs].

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.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, Ruibin Yuan, Haihong Wu, Hongquan Lin, Wenhao Huang, Jiajun Zhang, Wenhu Chen, Chenghua Lin, Jie Fu, Min Yang, Shiwen Ni, and Ge Zhang. COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning, March 2024a. URL http://arxiv.org/abs/2403.18058. arXiv:2403.18058 [cs] version: 1.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A Recipe for Long Context Alignment of Large Language Models, January 2024b. URL http://arxiv.org/abs/2401.18058. arXiv:2401.18058 [cs].

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed Movies: Story Based Retrieval with Contextual Embeddings, October 2020. URL http://arxiv.org/abs/2005.04208. arXiv:2005.04208 [cs].

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, May 2022. URL http://arxiv.org/abs/2104.00650. arXiv:2104.00650 [cs].

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, October 2018. URL http://arxiv.org/abs/1611.09268. arXiv:1611.09268 [cs].

Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue, June 2019. URL http://arxiv.org/abs/1906.07220. arXiv:1906.07220 [cs].

Jack Bandy and Nicholas Vincent. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. KsponSpeech: Korean Spontaneous Speech Corpus for Automatic Speech Recognition. *Applied Sciences*, 10(19):6936, January 2020. ISSN 2076-3417. doi: 10/gtwwck. URL https://www.mdpi.com/2076-3417/10/19/6936. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, Jordan Boyd-Graber, and Benjamin Van Durme. MegaWika: Millions of reports and their sources across 50 diverse languages, July 2023. URL http://arxiv.org/abs/2307.07049. arXiv:2307.07049 [cs].

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, December 2020. ISSN 2307-387X. doi: 10/gjzgwj. URL http://arxiv.org/abs/2002.00293. arXiv:2002.00293 [cs].

Timo Baumann, Arne Köhn, and Felix Hennig. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53(2):303–329, June 2019. ISSN 1574-0218. doi: 10/gq5xdf. URL https://doi.org/10.1007/s10579-017-9410-y.

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.

Anish Bhanushali, Grant Bridgman, Deekshitha G, Prasanta Ghosh, Pratik Kumar, Saurabh Kumar, Adithya Raj Kolladath, Nithya Ravi, Aaditeshwar Seth, Ashish Seth, Abhayjeet Singh, Vrunda Sukhadia, Umesh S, Sathvik Udupa, and Lodagala V. S. V. Durga Prasad. Gram Vaani ASR Challenge on spontaneous telephone speech recordings in regional variations of Hindi. In *Interspeech 2022*, pp. 3548–3552. ISCA, September 2022. doi: 10/gtsqzn. URL https://www.isca-archive.org/interspeech_2022/bhanushali22_interspeech.html.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages, August 2022. URL http://arxiv.org/abs/2208.12666. arXiv:2208.12666 [cs, eess].

Mariam Biltawi, Arafat Awajan, and Sara Tedmori. Arabic Reading Comprehension Benchmarks Created Semiautomatically. In *2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1–6, Giza, Egypt, November 2020. IEEE. ISBN 978-1-72818-855-3. doi: 10/g6k6b8. URL https://ieeexplore.ieee.org/document/9300111/.

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.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language, November 2019. URL http://arxiv.org/abs/1911.11641. arXiv:1911.11641 [cs].

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.

Marcely Zanon Boito, William N. Havard, Mahault Garnerin, Eric Le Ferrand, and Laurent Besacier. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible, February 2020. URL http://arxiv.org/abs/1907.12895. arXiv:1907.12895 [cs].

Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly Supervised Action Labeling in Videos Under Ordering Constraints, July 2014. URL http://arxiv.org/abs/1407.1208. arXiv:1407.1208 [cs].

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

Michael Boratko, Xiang Lorraine Li, Rajarshi Das, Tim O'Gorman, Dan Le, and Andrew McCallum. ProtoQA: A Question Answering Dataset for Prototypical Common-Sense Reasoning, October 2020. URL http://arxiv.org/abs/2005.00771. arXiv:2005.00771 [cs].

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. Learning To Split and Rephrase From Wikipedia Edit History, August 2018. URL http://arxiv.org/abs/1808.09468. arXiv:1808.09468 [cs].

Florian Boudin and Ygor Gallina. Redefining Absent Keyphrases and their Effect on Retrieval Effectiveness. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4185–4193, Online, June 2021. Association for

Computational Linguistics. doi: 10/g6k64d. URL https://aclanthology.org/2021.naacl-main.330.

A.R. Bradlow. ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings, 2010. URL https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar.

Alex Bravo, Janet Pinero, Nuria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1):55, February 2015. ISSN 1471-2105. doi: 10/f7kn8s. URL https://doi.org/10.1186/s12859-015-0472-9.

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., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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, July 2020b. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline, September 2017. URL http://arxiv.org/abs/1709.05522. arXiv:1709.05522 [cs].

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, February 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, February 2018b. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset, September 2019. URL http://arxiv.org/abs/1909.05358. arXiv:1909.05358 [cs].

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base, June 2022. URL http://arxiv.org/abs/2007.03875. arXiv:2007.03875 [cs].

Shuyang Cao and Lu Wang. Controllable Open-ended Question Generation with A New Question Type Ontology, June 2021. URL http://arxiv.org/abs/2107.00152. arXiv:2107.00152 [cs].

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI Meeting Corpus: A Pre-announcement. In Steve Renals and Samy Bengio (eds.), *Machine Learning for Multimodal Interaction*, volume 3869, pp. 28–39. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-32549-9. doi: 10.1007/11677482_3. URL http://link.springer.com/10.1007/11677482_3. Series Title: Lecture Notes in Computer Science.

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.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A Short Note about Kinetics-600, August 2018. URL http://arxiv.org/abs/1808.01340. arXiv:1808.01340 [cs].

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient Intent Detection with Dual Sentence Encoders. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah (eds.), *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, Online, July 2020. Association for Computational Linguistics. doi: 10/gjhzzs. URL https://aclanthology.org/2020.nlp4convai-1.5. arXiv:2003.04807 [cs].

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 2023. URL http://dx.doi.org/10.2139/ssrn.4789403.

Alberto Cetoli, Mohammad Akbari, Stefano Bragaglia, Andrew D. O'Harney, and Marc Sloan. Named Entity Disambiguation using Deep Learning on Graphs. volume 11438, pp. 78–86. 2019. doi: 10.1007/978-3-030-15719-7_10. URL http://arxiv.org/abs/1810.09164. arXiv:1810.09164 [cs].

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural Legal Judgment Prediction in English, June 2019a. URL http://arxiv.org/abs/1906.02059. arXiv:1906.02059 [cs].

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-Scale Multi-Label Text Classification on EU Legislation, June 2019b. URL http://arxiv.org/abs/1906.02192. arXiv:1906.02192 [cs].

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations, January 2021. URL http://arxiv.org/abs/2101.10726. arXiv:2101.10726 [cs].

Jun Shern Chan, Michael Pieler, Jonathan Jao, Jérémy Scheurer, and Ethan Perez. Few-shot Adaptation Works with UnpredicTable Data, August 2022. URL http://arxiv.org/abs/2208.01009. arXiv:2208.01009 [cs].

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.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. Hierarchical Pre-training for Sequence Labelling in Spoken Dialog, February 2021. URL http://arxiv.org/abs/2009.11152. arXiv:2009.11152 [cs].

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.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A Corpus of Campsite Negotiation Dialogues for Automatic Negotiation Systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3167–3185, Online, June 2021. Association for Computational Linguistics. doi: 10/gtsqxv. URL https://aclanthology.org/2021.naacl-main.254.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017, Online, June 2021a. Association for Computational Linguistics. doi: 10/gtsqxt. URL https://aclanthology.org/2021.naacl-main.239.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio, June 2021b. URL http://arxiv.org/abs/2106.06909. arXiv:2106.06909 [cs, eess].

Hannah Chen, Yangfeng Ji, and David Evans. Finding Friends and Flipping Frenemies: Automatic Paraphrase Dataset Augmentation Using Graph Theory, November 2020a. URL http://arxiv.org/abs/2011.01856. arXiv:2011.01856 [cs].

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions, June 2024. URL http://arxiv.org/abs/2406.04325. arXiv:2406.04325 [cs].

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. PLACES: Prompting Language Models for Social Conversation Synthesis, February 2023a. URL http://arxiv.org/abs/2302.03269. arXiv:2302.03269 [cs].

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. SummScreen: A Dataset for Abstractive Screenplay Summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8602–8615, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10/gtsqxx. URL https://aclanthology.org/2022.acl-long.589.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention, June 2019. URL http://arxiv.org/abs/1905.12866. arXiv:1905.12866 [cs].

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026–1036, Online, November 2020b. Association for Computational Linguistics. doi: 10/gpmd4x. URL https://aclanthology.org/2020.findings-emnlp.91.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A Theorem-driven Question Answering dataset, December 2023b. URL http://arxiv.org/abs/2305.12524. arXiv:2305.12524 [cs].

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

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5062–5074, Online, August 2021c. Association for Computational Linguistics. doi: 10/gtsqxs. URL https://aclanthology.org/2021.findings-acl.449.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. KETOD: Knowledge-Enriched Task-Oriented Dialogue, May 2022b. URL http://arxiv.org/abs/2205.05589. arXiv:2205.05589 [cs].

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation, March 2022. URL http://arxiv.org/abs/2108.06712. arXiv:2108.06712 [cs].

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. SalesBot: Transitioning from Chit-Chat to Task-Oriented Dialogues. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6143–6158, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10/gtsqxw. URL https://aclanthology.org/2022.acl-long.425.

Hasna Chouikhi, Manel Aloui, Cyrine Ben Hammou, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. GemmAr: Enhancing LLMs Through Arabic Instruction-Tuning, July 2024. URL http://arxiv.org/abs/2407.02147. arXiv:2407.02147 [cs].

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, February 2017. URL http://arxiv.org/abs/1703.00056. arXiv:1703.00056 [cs, stat].

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 729–738, November 2019. doi: 10/gkz233. URL http://arxiv.org/abs/1910.03262. arXiv:1910.03262 [cs].

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. HAA500: Human-Centric Atomic Action Dataset with Curated Videos, August 2021. URL http://arxiv.org/abs/2009.05224. arXiv:2009.05224 [cs, eess].

Christopher Cieri, David Miller, and Kevin Walker. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions, May 2019. URL http://arxiv.org/abs/1905.10044. arXiv:1905.10044 [cs].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL http://arxiv.org/abs/1803.05457. arXiv:1803.05457 [cs].

Steven Coats. A Corpus of Regional American Language from YouTube. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, 2019. URL https://www.semanticscholar.org/paper/A-Corpus-of-Regional-American-Language-from-YouTube-Coats/bc428db824d261794a7e081a53c4315b8e02f855.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

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

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech, May 2022. URL http://arxiv.org/abs/2205.12446. arXiv:2205.12446 [cs, eess] version: 1.

Sam Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations, July 2020. URL http://arxiv.org/abs/2005.08866. arXiv:2005.08866 [cs].

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces, December 2018. URL http://arxiv.org/abs/1805.10190. arXiv:1805.10190 [cs].

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.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset, July 2018. URL http://arxiv.org/abs/1804.02748. arXiv:1804.02748 [cs].

Verna Dankers, Christopher Lucas, and Ivan Titov. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3608–3626, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10/g6k6xn. URL https://aclanthology.org/2022.acl-long.252.

Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2634–2641, Portland, OR, USA, June 2013. IEEE. ISBN 978-0-7695-4989-7. doi: 10/gtsqzr. URL http://ieeexplore.ieee.org/document/6619184/.

Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota Smarthome: Real-World Activities of Daily Living. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 833–842, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10/ghfjc7. URL https://ieeexplore.ieee.org/document/9008135/.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning, September 2019. URL http://arxiv.org/abs/1908.05803. arXiv:1908.05803 [cs].

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.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. Earnings-22: A Practical Benchmark for Accents in the Wild, March 2022. URL http://arxiv.org/abs/2203.15591. arXiv:2203.15591 [cs].

Ugur Demir, Yogesh S. Rawat, and Mubarak Shah. TinyVIRAT: Low-resolution Video Action Recognition, July 2020. URL http://arxiv.org/abs/2007.07355. arXiv:2007.07355 [cs, eess].

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. Open-source Multispeaker Corpora of the English Accents in the British Isles. 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. 6532–6541, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.804.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a Generalist Agent for the Web, December 2023. URL http://arxiv.org/abs/2306.06070. arXiv:2306.06070 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023. URL http://arxiv.org/abs/2305.14314. arXiv:2305.14314 [cs].

Ashwin Devaraj, Iain J. Marshall, Byron C. Wallace, and Junyi Jessy Li. Paragraph-level Simplification of Medical Texts, April 2021. URL http://arxiv.org/abs/2104.05767. arXiv:2104.05767 [cs].

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. MS2: Multi-Document Summarization of Medical Studies, November 2021. URL http://arxiv.org/abs/2104.06486. arXiv:2104.06486 [cs].

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10/gtsqzk. URL https://aclanthology.org/N19-1202.

Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large Scale Holistic Video Understanding, December 2020. URL http://arxiv.org/abs/1904.11451. arXiv:1904.11451 [cs].

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims, January 2021. URL http://arxiv.org/abs/2012.00614. arXiv:2012.00614 [cs].

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The Second Conversational Intelligence Challenge (ConvAI2), January 2019a. URL http://arxiv.org/abs/1902.00098. arXiv:1902.00098 [cs].

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-Powered Conversational agents, February 2019b. URL http://arxiv.org/abs/1811.01241. arXiv:1811.01241 [cs].

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations, May 2023. URL http://arxiv.org/abs/2305.14233. arXiv:2305.14233 [cs].

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.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using Radio Archives for Low-Resource Speech Recognition: Towards an Intelligent Virtual Assistant for Illiterate Users, April 2021. URL http://arxiv.org/abs/2104.13083. arXiv:2104.13083 [cs].

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale, September 2018. URL http://arxiv.org/abs/1808.10583. arXiv:1808.10583 [cs].

Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced Webly-supervised Learning for Video Recognition, August 2020. URL http://arxiv.org/abs/2003.13042. arXiv:2003.13042 [cs].

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. SANAD: Single-label Arabic News Articles Dataset for automatic text categorization. *Data in Brief*, 25:104076, August 2019. ISSN 23523409. doi: 10/g6k6cn. URL https://linkinghub.elsevier.com/retrieve/pii/S2352340919304305.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (eds.), *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 207–219, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10/gtsqx4. URL https://aclanthology.org/W17-5526.

Ibrahim Abu El-khair. 1.5 billion words Arabic Corpus, November 2016. URL http://arxiv.org/abs/1611.04033. arXiv:1611.04033 [cs].

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.

Ashraf Elnagar and Omar Einea. BRAD 1.0: Book reviews in Arabic dataset. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pp. 1–8, Agadir, Morocco, November 2016. IEEE. ISBN 978-1-5090-4320-0. doi: 10/g6k6jm. URL http://ieeexplore.ieee.org/document/7945800/.

Hady ElSahar and Samhaa R. El-Beltagy. Building Large Arabic Multi-domain Resources for Sentiment Analysis. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 23–34, Cham, 2015. Springer International Publishing. ISBN 978-3-319-18117-2. doi: 10/g6k58r.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences, December 2020. URL http://arxiv.org/abs/2012.15738. arXiv:2012.15738 [cs].

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! Predicting Unintentional Action in Video. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 916–926, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10/ghbckh. URL https://ieeexplore.ieee.org/document/9156404/.

Mihail Eric and Christopher D. Manning. Key-Value Retrieval Networks for Task-Oriented Dialogue, July 2017. URL http://arxiv.org/abs/1705.05414. arXiv:1705.05414 [cs].

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines, December 2019. URL http://arxiv.org/abs/1907.01669. arXiv:1907.01669 [cs].

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.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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. 6866–6880, Online, August 2021. Association for Computational Linguistics. doi: 10/gmf9qs. URL https://aclanthology.org/2021.acl-long.535.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model, June 2019. URL http://arxiv.org/abs/1906.01749. arXiv:1906.01749 [cs].

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.

Manaal Faruqui and Dipanjan Das. Identifying Well-formed Natural Language Questions, August 2018. URL http://arxiv.org/abs/1808.09419. arXiv:1808.09419 [cs].

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. TWEETSUMM – A Dialog Summarization Dataset for Customer Service, November 2021. URL http://arxiv.org/abs/2111.11894. arXiv:2111.11894 [cs].

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6162–6176, 2021. doi: 10/g6k938. URL http://arxiv.org/abs/2109.12595. arXiv:2109.12595 [cs].

David F. Fouhey, Wei-cheng Kuo, Alexei A. Efros, and Jitendra Malik. From Lifestyle Vlogs to Everyday Interactions, December 2017. URL http://arxiv.org/abs/1712.02310. arXiv:1712.02310 [cs].

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario, August 2021. URL http://arxiv.org/abs/2104.03603. arXiv:2104.03603 [cs, eess].

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.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. Airavata: Introducing Hindi Instruction-tuned LLM, February 2024. URL http://arxiv.org/abs/2401.15006. arXiv:2401.15006 [cs].

Ygor Gallina, Florian Boudin, and Beatrice Daille. KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura (eds.), *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 130–135, Tokyo, Japan, October 2019. Association for Computational Linguistics. doi: 10/g6k64k. URL https://aclanthology.org/W19-8617.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage, November 2021. URL http://arxiv.org/abs/2111.09344. arXiv:2111.09344 [cs, stat].

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In Chu-Ren Huang and Dan Jurafsky (eds.), *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 340–348, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1039.

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.

Garofolo, John S., Lamel, Lori F., Fisher, William M., Pallett, David S., Dahlgren, Nancy L., Zue, Victor, and Fiscus, Jonathan G. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993. URL https://catalog.ldc.upenn.edu/LDC93S1. Artwork Size: 715776 KB Pages: 715776 KB.

Murilo Gleyson Gazzola, Sidney Evaldo Leal, and Sandra Maria Aluísio. Predição da complexidade textual de recursos educacionais abertos em português. *Symposium in Information and Human Language Technology - STIL*, 2019. URL https://repositorio.usp.br/item/002971271.

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.

Elizabeth Jasmi George and Radhika Mamidi. Conversational implicatures in English dialogue: Annotated dataset, November 2019. URL http://arxiv.org/abs/1911.10704. arXiv:1911.10704 [cs].

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. Multilingual and Cross-Lingual Intent Detection from Spoken Data, April 2021. URL http://arxiv.org/abs/2104.08524. arXiv:2104.08524 [cs].

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion, March 2019. URL http://arxiv.org/abs/1902.10526. arXiv:1902.10526 [cs].

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10/gmjqgr. URL https://aclanthology.org/D19-5409.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 517–520 vol.1, San Francisco, CA, USA, 1992. IEEE. ISBN 978-0-7803-0532-8. doi: 10/fp48kw. URL http://ieeexplore.ieee.org/document/225858/.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. RumourEval 2019: Determining Rumour Veracity and Support for Rumours, September 2018. URL http://arxiv.org/abs/1809.06683. arXiv:1809.06683 [cs].

Tear Gosling, Alpin Dale, and Yinhe Zheng. PIPPA: A Partially Synthetic Conversational Dataset, August 2023. URL http://arxiv.org/abs/2308.05884. arXiv:2308.05884 [cs].

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.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, June 2017. URL http://arxiv.org/abs/1706.04261. arXiv:1706.04261 [cs].

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video, March 2022. URL http://arxiv.org/abs/2110.07058. arXiv:2110.07058 [cs].

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions, April 2018. URL http://arxiv.org/abs/1705.08421. arXiv:1705.08421 [cs].

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In *Proceedings of the Web Conference 2021*, pp. 3477–3488, April 2021. doi: 10/gnnfbt. URL http://arxiv.org/abs/2011.07743. arXiv:2011.07743 [cs].

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, January 2022. ISSN 2691-1957, 2637-8051. doi: 10/gnmkjx. URL http://arxiv.org/abs/2007.15779. arXiv:2007.15779 [cs].

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, and Xiangang Li. DiDiSpeech: A Large Scale Mandarin Speech Corpus, February 2021. URL http://arxiv.org/abs/2010.09275. arXiv:2010.09275 [eess].

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. Disfl-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering, June 2021. URL http://arxiv.org/abs/2106.04016. arXiv:2106.04016 [cs].

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018a. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1440.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. Semantic Parsing for Task Oriented Dialog using Hierarchical Representations, October 2018b. URL http://arxiv.org/abs/1810.07942. arXiv:1810.07942 [cs].

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating Summaries from User Videos. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, volume 8695, pp. 505–520. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10583-3. doi: 10.1007/978-3-319-10584-0_33. URL http://link.springer.com/10.1007/978-3-319-10584-0_33. Series Title: Lecture Notes in Computer Science.

David Ha and Douglas Eck. A Neural Representation of Sketch Drawings, May 2017. URL http://arxiv.org/abs/1704.03477. arXiv:1704.03477 [cs, stat].

Hessel Haagsma, Johan Bos, and Malvina Nissim. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. 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. 279–287, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.35.

Amirhossein Habibian, Thomas Mensink, and Cees G.M. Snoek. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 17–26, Orlando Florida USA, November 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10/ggs25n. URL https://dl.acm.org/doi/10.1145/2647868.2654913.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data, October 2023. URL http://arxiv.org/abs/2304.08247. arXiv:2304.08247 [cs].

Moshe Hazoom, Vibhor Malik, and Ben Bogin. Text-to-SQL in the Wild: A Naturally-Occurring Dataset Based on Stack Exchange Data, June 2021. URL http://arxiv.org/abs/2106.05006. arXiv:2106.05006 [cs].

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling Strategy and Generation in Negotiation Dialogues, August 2018. URL http://arxiv.org/abs/1808.09637. arXiv:1808.09637 [cs].

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10/gfsvdw. URL http://ieeexplore.ieee.org/document/7298698/.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. doi: 10/cz3442. URL https://aclanthology.org/H90-1021.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, November 2022. URL http://arxiv.org/abs/2207.00220. arXiv:2207.00220 [cs].

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.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video with Temporal Language. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1380–1390, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10/gtsqzq. URL https://aclanthology.org/D18-1168.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review, November 2021. URL http://arxiv.org/abs/2103.06268. arXiv:2103.06268 [cs].

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.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *Lecture Notes in Computer Science*, volume 11096, pp. 198–208. Springer, Cham, 2018. doi: 10.1007/978-3-319-99579-3_21. URL http://arxiv.org/abs/1805.04699. arXiv:1805.04699 [cs].

Carlos Daniel Hernandez Mena, David Erik Mollberg, Michal Borsky, and Jon Gudnason. Samrómur Children: An Icelandic Speech Corpus. In Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Helene Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 995–1002, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.105.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, September 2020. URL http://arxiv.org/abs/2003.11080. arXiv:2003.11080 [cs].

Albert Huang. 'Tis but Thy Name: Semantic Question Answering Evaluation with 11M Names for 1M Entities, February 2022. URL http://arxiv.org/abs/2202.13581. arXiv:2202.13581 [cs].

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning, September 2019. URL http://arxiv.org/abs/1909.00277. arXiv:1909.00277 [cs].

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient Attentions for Long Document Summarization, April 2021. URL http://arxiv.org/abs/2104.02112. arXiv:2104.02112 [cs].

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A Holistic Dataset for Movie Understanding, July 2020. URL http://arxiv.org/abs/2007.10937. arXiv:2007.10937 [cs].

Moustafa Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A Hierarchical Deep Temporal Model for Group Activity Recognition, April 2016. URL http://arxiv.org/abs/1511.06040. arXiv:1511.06040 [cs].

Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS Challenge on Action Recognition for Videos "in the Wild". *Computer Vision and Image Understanding*, 155:1–23, February 2017. ISSN 10773142. doi: 10/f9rwnr. URL http://arxiv.org/abs/1604.06182. arXiv:1604.06182 [cs].

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, December 2020. ISSN 1549-9596. doi: 10/gmjg8b. URL https://doi.org/10.1021/acs.jcim.0c00675. Publisher: American Chemical Society.

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

Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient Long-Text Understanding with Short-Text Models, December 2022. URL http://arxiv.org/abs/2208.00748. arXiv:2208.00748 [cs].

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. Learning a Neural Semantic Parser from User Feedback, April 2017. URL http://arxiv.org/abs/1704.08760. arXiv:1704.08760 [cs].

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. Search-based Neural Structured Learning for Sequential Question Answering. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1821–1831, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10/gf6nx8. URL https://aclanthology.org/P17-1167.

Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian languages, December 2022. URL http://arxiv.org/abs/2208.11761. arXiv:2208.11761 [cs, eess].

Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. LEMMA: A Multi-view Dataset for Learning Multi-agent Multi-task Activities, July 2020. URL http://arxiv.org/abs/2007.15781. arXiv:2007.15781 [cs].

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF Model for Sentence Alignment in Text Simplification, August 2021. URL http://arxiv.org/abs/2005.02324. arXiv:2005.02324 [cs].

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification, November 2020. URL http://arxiv.org/abs/2011.03088. arXiv:2011.03088 [cs].

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams, September 2020. URL http://arxiv.org/abs/2009.13081. arXiv:2009.13081 [cs].

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering, September 2019. URL http://arxiv.org/abs/1909.06146. arXiv:1909.06146 [cs, q-bio].

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, May 2017. URL http://arxiv.org/abs/1705.03551. arXiv:1705.03551 [cs].

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.

Juraj Juraska, Kevin K. Bowden, and Marilyn Walker. ViGGO: A Video Game Corpus for Data-To-Text Generation in Open-Domain Conversation, October 2019. URL http://arxiv.org/abs/1910.12129. arXiv:1910.12129 [cs].

Tomasz Jurczyk, Michael Zhai, and Jinho D. Choi. SelQA: A New Benchmark for Selection-based Question Answering, October 2016. URL http://arxiv.org/abs/1606.08513. arXiv:1606.08513 [cs].

Aditya Kanade, Petros Maniatis, Gogul Balakrishnan, and Kensen Shi. Learning and Evaluating Contextual Embedding of Source Code, August 2020. URL http://arxiv.org/abs/2001.00059. arXiv:2001.00059 [cs].

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.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5. doi: 10/gf4hdn. URL https://ieeexplore.ieee.org/document/6909619.

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. Golos: Russian Dataset for Speech Research, June 2021. URL http://arxiv.org/abs/2106.10161. arXiv:2106.10161 [eess].

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data, February 2020. URL http://arxiv.org/abs/1909.12434. arXiv:1909.12434 [cs, stat].

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, May 2017. URL http://arxiv.org/abs/1705.06950. arXiv:1705.06950 [cs].

Yerbolat Khassanov, Saida Mussakhojayeva, Almas Mirzakhmetov, Alen Adiyev, Mukhamet Nurpeiissov, and Huseyin Atakan Varol. A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline, January 2021. URL http://arxiv.org/abs/2009.10334. arXiv:2009.10334 [cs, eess].

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTaiL: A Textual Entailment Dataset from Science Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10/grm22d. URL https://ojs.aaai.org/index.php/AAAI/article/view/12022.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A Dataset for Question Answering via Sentence Composition, February 2020. URL http://arxiv.org/abs/1910.11473. arXiv:1910.11473 [cs].

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. ProsocialDialog: A Prosocial Backbone for Conversational Agents, October 2022. URL http://arxiv.org/abs/2205.12688. arXiv:2205.12688 [cs].

Seonhoon Kim, Inho Kang, and Nojun Kwak. Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information, November 2018. URL http://arxiv.org/abs/1805.11360. arXiv:1805.11360 [cs].

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models, March 2024. URL http://arxiv.org/abs/2310.08491. arXiv:2310.08491 [cs].

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. FT Speech: Danish Parliament Speech Corpus, October 2020. URL http://arxiv.org/abs/2005.12368. arXiv:2005.12368 [cs, eess].

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pp. 52–55. ISCA, August 2018. doi: 10/gtwwbs. URL https://www.isca-archive.org/sltu_2018/kjartansson18_sltu.html.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In Dorothee Beermann, Laurent Besacier, Sakriani Sakti, and Claudia Soria (eds.), *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pp. 21–27, Marseille, France, May 2020. European Language Resources association. ISBN 979-10-95546-35-1. URL https://aclanthology.org/2020.sltu-1.3.

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

Rostislav Kolobov, Olga Okhapkina, Olga Omelchishina, Andrey Platunov, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. MediaSpeech: Multilanguage ASR Benchmark and Dataset, March 2021. URL http://arxiv.org/abs/2103.16193. arXiv:2103.16193 [cs, eess].

Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-Augmented Dialogue Generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8460–8478, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10/gr75db. URL https://aclanthology.org/2022.acl-long.579.

Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8657–8666, October 2019. doi: 10/ghfhxx. URL https://ieeexplore.ieee.org/document/9009579. ISSN: 2380-7504.

Anastassia Kornilova and Vlad Eidelman. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 48–56, 2019. doi: 10/gtwwtd. URL http://arxiv.org/abs/1910.00523. arXiv:1910.00523 [cs].

Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 4423–4428, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/535_Paper.pdf.

Danijel Korzinek, Krzysztof Marasek, Lukasz Brocki, and Krzysztof Wolk. Polish Read Speech Corpus for Speech Tools and Services, June 2017. URL http://arxiv.org/abs/1706.00245. arXiv:1706.00245 [cs].

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*, 2023.

Jonáš Kratochvil, Peter Polák, and Ondřej Bojar. Large Corpus of Czech Parliament Plenary Hearings. 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. 6363–6367, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.781.

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.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pp. 2556–2563, Barcelona, Spain, November 2011. IEEE. ISBN 978-1-4577-1102-2. doi: 10/fxpf8k. URL http://ieeexplore.ieee.org/document/6126543/.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5. doi: 10/gqdc3v. URL https://ieeexplore.ieee.org/document/6909500/.

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.

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7(1):281, August 2020. ISSN 2052-4463. doi: 10/gr4ftn. URL https://www.nature.com/articles/s41597-020-00620-0. Publisher: Nature Publishing Group.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant Conversations – Democratizing Large Language Model Alignment, October 2023. URL http://arxiv.org/abs/2304.07327. arXiv:2304.07327 [cs].

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations, December 2017. URL http://arxiv.org/abs/1704.04683. arXiv:1704.04683 [cs].

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks, June 2018. URL http://arxiv.org/abs/1711.00350. arXiv:1711.00350 [cs].

Lander, T. CSLU: 22 Languages Corpus, November 2005. URL https://catalog.ldc.upenn.edu/LDC2005S26.

Lander, T. CSLU: Foreign Accented English Release 1.2, May 2007. URL https://catalog.ldc.upenn.edu/LDC2007S08. Artwork Size: 1468006 KB Pages: 1468006 KB.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10/gqrfvv. URL https://aclanthology.org/D19-1131.

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.

Remi Lebret, David Grangier, and Michael Auli. Neural Text Generation from Structured Data with Application to the Biography Domain, September 2016. URL http://arxiv.org/abs/1603.07771. arXiv:1603.07771 [cs].

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.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks, October 2022. URL http://arxiv.org/abs/2210.14712. arXiv:2210.14712 [cs].

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or No Deal? End-to-End Learning for Negotiation Dialogues, June 2017. URL http://arxiv.org/abs/1706.05125. arXiv:1706.05125 [cs].

Guanyu Li, Hongzhi Yu, Thomas Fang Zheng, Jinghao Yan, and Shipeng Xu. Free linguistic and speech resources for Tibetan. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 733–736, Kuala Lumpur, December 2017. IEEE. ISBN 978-1-5386-1542-3. doi: 10/gtsqzh. URL http://ieeexplore.ieee.org/document/8282130/.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-X: Multilingual Replicable Instruction-Following Models with Low-Rank Adaptation, October 2023a. URL http://arxiv.org/abs/2305.15011. arXiv:2305.15011 [cs].

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. MTOP: A Comprehensive Multilingual Task-Oriented Semantic Parsing Benchmark. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2950–2962, Online, April 2021a. Association for Computational Linguistics. doi: 10/gtsqxq. URL https://aclanthology.org/2021.eacl-main.257.

Hongyu Li, Seohyun Kim, and Satish Chandra. Neural Code Search Evaluation Dataset, October 2019a. URL http://arxiv.org/abs/1908.09804. arXiv:1908.09804 [cs].

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark, May 2024. URL http://arxiv.org/abs/2311.17005. arXiv:2311.17005 [cs].

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, 2019b. doi: 10.1109/TETCI.2019.2892755.

Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles, August 2021b. URL http://arxiv.org/abs/2104.00946. arXiv:2104.00946 [cs].

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, 2023b.

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, Taipei, Taiwan, December 2023c. IEEE. ISBN 979-8-3503-0689-7. doi: 10/gtsqzc. URL https://ieeexplore.ieee.org/document/10389689/.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. End-to-End Trainable Non-Collaborative Dialog System, November 2019c. URL http://arxiv.org/abs/1911.10742. arXiv:1911.10742 [cs].

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-Level Code Generation with AlphaCode. *Science*, 378(6624):1092–1097, December 2022. ISSN 0036-8075, 1095-9203. doi: 10/grggxf. URL http://arxiv.org/abs/2203.07814. arXiv:2203.07814 [cs].

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A New Dataset and Benchmark on Animated GIF Description, April 2016. URL http://arxiv.org/abs/1604.02748. arXiv:1604.02748 [cs].

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge, June 2023d. URL http://arxiv.org/abs/2303.14070. arXiv:2303.14070 [cs].

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step, May 2023. URL http://arxiv.org/abs/2305.20050. arXiv:2305.20050 [cs].

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models, September 2020a. URL http://arxiv.org/abs/2005.00683. arXiv:2005.00683 [cs].

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning, November 2020b. URL http://arxiv.org/abs/1911.03705. arXiv:1911.03705 [cs].

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge, July 2021a. URL http://arxiv.org/abs/2101.00376. arXiv:2101.00376 [cs].

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. Reasoning Over Paragraph Effects in Situations, December 2019. URL http://arxiv.org/abs/1908.05852. arXiv:1908.05852 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling, June 2021b. URL http://arxiv.org/abs/2106.02787. arXiv:2106.02787 [cs].

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems, October 2017. URL http://arxiv.org/abs/1705.04146. arXiv:1705.04146 [cs].

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding, March 2017. URL http://arxiv.org/abs/1703.07475. arXiv:1703.07475 [cs].

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.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8386–8390, Vancouver, BC, Canada, May 2013. IEEE. ISBN 978-1-4799-0356-6. doi: 10/gtsqxn. URL http://ieeexplore.ieee.org/document/6639301/.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. VIOLIN: A Large-Scale Dataset for Video-and-Language Inference, March 2020. URL http://arxiv.org/abs/2003.11618. arXiv:2003.11618 [cs].

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.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning, April 2024b. URL http://arxiv.org/abs/2312.15685. arXiv:2312.15685 [cs].

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents, October 2023c. URL http://arxiv.org/abs/2308.03688. arXiv:2308.03688 [cs].

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking Natural Language Understanding Services for building Conversational Agents, March 2019. URL http://arxiv.org/abs/1903.05566. arXiv:1903.05566 [cs].

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

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4335–4347, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10/gtsqxr. URL https://aclanthology.org/2021.emnlp-main.356.

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.

Annie Louis, Dan Roth, and Filip Radlinski. "I'd rather just go to bed": Understanding Indirect Answers, October 2020. URL http://arxiv.org/abs/2010.03450. arXiv:2010.03450 [cs].

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.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems, February 2016. URL http://arxiv.org/abs/1506.08909. arXiv:1506.08909 [cs].

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering, October 2022. URL http://arxiv.org/abs/2209.09513. arXiv:2209.09513 [cs].

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.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria Everyday Activities Dataset, February 2024. URL http://arxiv.org/abs/2402.13349. arXiv:2402.13349 [cs].

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems, October 2023. URL http://arxiv.org/abs/2305.14536. arXiv:2305.14536 [cs].

Kikuo Maekawa. Corpus of spontaneous Japanese: its design and evaluation. In *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. paper MMO2, 2003. URL https://www.isca-archive.org/sspr_2003/maekawa03_sspr.html.

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.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-Curated Questions and Attributed Answers, April 2024. URL http://arxiv.org/abs/2309.07852. arXiv:2309.07852 [cs].

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

Srikanth Malla, Behzad Dariush, and Chiho Choi. TITAN: Future Forecast Using Action Priors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11183–11193, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10/gg99rg. URL https://ieeexplore.ieee.org/document/9156550/.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts, July 2013. URL http://arxiv.org/abs/1307.5336. arXiv:1307.5336 [cs, q-fin].

Ikram Mamtimin, Wenqiang Du, and Askar Hamdulla. M2ASR-KIRGHIZ: A Free Kirghiz Speech Database and Accompanied Baselines. *Information*, 14(1):55, January 2023. ISSN 2078-2489. doi: 10/gtsqzm. URL https://www.mdpi.com/2078-2489/14/1/55.

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.

Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10/d5bs7p. URL https://ieeexplore.ieee.org/document/5206557/.

Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. Event Representations for Automated Story Generation with Deep Neural Nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10/g6k72p. URL http://arxiv.org/abs/1706.01331. arXiv:1706.01331 [cs].

Scott Martin, Shivani Poddar, and Kartikeya Upasani. MuDoCo: Corpus for Multidomain Coreference Resolution and Referring Expression Generation. 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. 104–111, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.13.

Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2874–2882, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72815-023-9. doi: 10/gh5k47. URL https://ieeexplore.ieee.org/document/9022297/.

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.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models, September 2016. URL http://arxiv.org/abs/1609.07843. arXiv:1609.07843 [cs].

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, July 2019. URL http://arxiv.org/abs/1906.03327. arXiv:1906.03327 [cs].

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. RareAct: A video dataset of unusual interactions, August 2020. URL http://arxiv.org/abs/2008.01018. arXiv:2008.01018 [cs].

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering, September 2018. URL http://arxiv.org/abs/1809.02789. arXiv:1809.02789 [cs].

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.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. Lila: A Unified Benchmark for Mathematical Reasoning, March 2023. URL http://arxiv.org/abs/2210.17517. arXiv:2210.17517 [cs].

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.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-Math: Unlocking the potential of SLMs in Grade School Math, February 2024. URL http://arxiv.org/abs/2402.14830. arXiv:2402.14830 [cs].

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.

David Erik Mollberg, Olafur Helgi Jonsson, Sunneva Thorsteinsdottir, Steinthor Steingremsson, Eydis Huld Magnusdottir, and Jon Gudnason. Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3463–3467, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.425.

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, 2019a.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and Aude Oliva. Moments in Time Dataset: one million videos for event understanding, February 2019b. URL http://arxiv.org/abs/1801.03150. arXiv:1801.03150 [cs].

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 2021a. URL http://arxiv.org/abs/2105.04489. arXiv:2105.04489 [cs, eess].

Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A. McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding, September 2021b. URL http://arxiv.org/abs/1911.00232. arXiv:1911.00232 [cs, eess].

Nuno Moniz and Luís Torgo. Multi-Source Social Feedback of Online News Feeds, January 2018. URL http://arxiv.org/abs/1801.07055. arXiv:1801.07055 [cs].

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 845–854, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10/ggt4wm. URL https://aclanthology.org/P19-1081.

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.

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. STAR: A Schema-Guided Dialog Dataset for Transfer Learning, October 2020. URL http://arxiv.org/abs/2010.11853. arXiv:2010.11853 [cs].

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. GLUCOSE: GeneraLized and COntextualized Story Explanations, October 2020. URL http://arxiv.org/abs/2009.07758. arXiv:2009.07758 [cs].

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.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. Neural Arabic Question Answering, June 2019. URL http://arxiv.org/abs/1906.05394. arXiv:1906.05394 [cs].

Nikola Mrksic, Diarmuid O Seaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural Belief Tracker: Data-Driven Dialogue State Tracking, April 2017. URL http://arxiv.org/abs/1606.03777. arXiv:1606.03777 [cs].

Nikola Mrkšić and Ivan Vulić. Fully Statistical Neural Belief Tracking, May 2018. URL http://arxiv.org/abs/1805.11350. arXiv:1805.11350 [cs].

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. QASR: QCRI Aljazeera Speech Resource – A Large Scale Annotated Arabic Speech Corpus, June 2021. URL http://arxiv.org/abs/2106.13000. arXiv:2106.13000 [cs, eess].

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.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10893–10906, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10/gtsqxz. URL https://aclanthology.org/2022.emnlp-main.748.

Will Myers, Tyler Etchart, and Nancy Fulda. Conversational Scaffolding: An Analogy-based Approach to Response Prioritization in Open-domain Dialogs:. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pp. 69–78, Valletta, Malta, 2020. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-395-7. doi: 10/gtsq86. URL http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0008939900690078.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: a large-scale speaker identification dataset, May 2018. URL http://arxiv.org/abs/1706.08612. arXiv:1706.08612 [cs].

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation, August 2024. URL http://arxiv.org/abs/2407.02371. arXiv:2407.02371 [cs].

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. DART: Open-Domain Structured Data Record to Text Generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 432–447, Online, June 2021. Association for Computational Linguistics. doi: 10/gnh49f. URL https://aclanthology.org/2021.naacl-main.37. arXiv:2007.02871 [cs].

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022. doi: 10/gtsqx3. URL https://aclanthology.org/2022.tacl-1.3. Place: Cambridge, MA Publisher: MIT Press.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization, August 2018. URL http://arxiv.org/abs/1808.08745. arXiv:1808.08745 [cs].

Linh The Nguyen, Nguyen Luong Tran, Long Doan, Manh Luong, and Dat Quoc Nguyen. A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation, August 2022. URL http://arxiv.org/abs/2208.04243. arXiv:2208.04243 [cs].

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. Advanced Semantics for Commonsense Knowledge Extraction. In *Proceedings of the Web Conference 2021*, pp. 2636–2647, April 2021. doi: 10/gnnffn. URL http://arxiv.org/abs/2011.00905. arXiv:2011.00905 [cs].

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. SeaLLMs – Large Language Models for Southeast Asia, December 2023. URL http://arxiv.org/abs/2312.00738. arXiv:2312.00738 [cs].

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding, May 2020. URL http://arxiv.org/abs/1910.14599. arXiv:1910.14599 [cs].

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E Dataset: New Challenges For End-to-End Generation, July 2017. URL http://arxiv.org/abs/1706.09254. arXiv:1706.09254 [cs].

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR, September 2023. URL http://arxiv.org/abs/2310.00274. arXiv:2310.00274 [cs].

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

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. QuerYD: A video dataset with high-quality text and audio narrations, February 2021. URL http://arxiv.org/abs/2011.11071. arXiv:2011.11071 [cs].

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition, April 2021. URL http://arxiv.org/abs/2104.02014. arXiv:2104.02014 [cs, eess].

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/.

Mariam Orabi, Hozayfa El Rifai, and Ashraf Elnagar. Classical Arabic Poetry: Classification based on Era. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–6, Antalya, Turkey, November 2020. IEEE. ISBN 978-1-72818-577-4. doi: 10/g6k6b2. URL https://ieeexplore.ieee.org/document/9316520/.

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.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color, October 2021. URL http://arxiv.org/abs/2110.08182. arXiv:2110.08182 [cs].

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B. Melton. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. *AMIA Annual Symposium Proceedings*, 2010:572–576, 2010. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041430/.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria Digital Twin: A New Benchmark Dataset for Egocentric 3D Machine Perception, June 2023. URL http://arxiv.org/abs/2306.06362. arXiv:2306.06362 [cs].

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, South Brisbane, Queensland, Australia, April 2015. IEEE. ISBN 978-1-4673-6997-8. doi: 10/gfv84w. URL http://ieeexplore.ieee.org/document/7178964/.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, June 2005. URL http://arxiv.org/abs/cs/0506075. arXiv:cs/0506075.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A Controlled Table-To-Text Generation Dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, Online, November 2020. Association for Computational Linguistics. doi: 10/gm3nmg. URL https://aclanthology.org/2020.emnlp-main.89.

Jeongkyun Park, Jung-Wook Hwang, Kwanghee Choi, Seung-Hyun Lee, Jun Hwan Ahn, Rae-Hong Park, and Hyung-Min Park. OLKAVS: An Open Large-Scale Korean Audio-Visual Speech Dataset, January 2023. URL http://arxiv.org/abs/2301.06375. arXiv:2301.06375 [cs].

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.

Panupong Pasupat and Percy Liang. Compositional Semantic Parsing on Semi-Structured Tables. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10/gfz98s. URL https://aclanthology.org/P15-1142.

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.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription, October 2021. URL http://arxiv.org/abs/2107.01091. arXiv:2107.01091 [cs, eess].

Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, June 2007. ISSN 1532-0464. doi: 10/fghjwr. URL https://www.sciencedirect.com/science/article/pii/S1532046406000645.

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.

F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *2016 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 724–732, June 2016. doi: 10/ggdmmw. URL https://ieeexplore.ieee.org/document/7780454. ISSN: 1063-6919.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. Generating Summaries with Topic Templates and Structured Convolutional Decoders, June 2019. URL http://arxiv.org/abs/1906.04687. arXiv:1906.04687 [cs].

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies toward Curating and Annotating Large Scale Dialogue Data. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4526–4536, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10/gkr9hj. URL https://aclanthology.org/D19-1460.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases?, September 2019. URL http://arxiv.org/abs/1909.01066. arXiv:1909.01066 [cs].

Thang M. Pham, Seunghyun Yoon, Trung Bui, and Anh Nguyen. PiC: A Phrase-in-Context Dataset for Phrase Understanding and Semantic Search, February 2023. URL http://arxiv.org/abs/2207.09068. arXiv:2207.09068 [cs].

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. BiMediX: Bilingual Medical Mixture of Experts LLM, February 2024. URL http://arxiv.org/abs/2402.13253. arXiv:2402.13253 [cs].

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. SDS-200: A Swiss German Speech to Standard German Text Corpus, May 2022. URL http://arxiv.org/abs/2205.09501. arXiv:2205.09501 [cs].

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, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pp. 2757–2761, October 2020. doi: 10/grk6mp. URL http://arxiv.org/abs/2012.03411. arXiv:2012.03411 [cs, eess].

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.

Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. Multilingual Event Linking to Wikidata, July 2022. URL http://arxiv.org/abs/2204.06535. arXiv:2204.06535 [cs].

Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. Database Search Results Disambiguation for Task-Oriented Dialog Systems. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1158–1173, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10/gtsqx2. URL https://aclanthology.org/2022.naacl-main.85.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4547–4557, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10/gk3btq. URL https://aclanthology.org/D19-1462.

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.

Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home Action Genome: Cooperative Compositional Action Understanding, May 2021. URL http://arxiv.org/abs/2105.05226. arXiv:2105.05226 [cs].

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain Yourself! Leveraging Language Models for Commonsense Reasoning, June 2019. URL http://arxiv.org/abs/1906.02361. arXiv:1906.02361 [cs].

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016a.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text, October 2016b. URL http://arxiv.org/abs/1606.05250. arXiv:1606.05250 [cs].

Kavitha Raju, Anjaly V, Ryan Lish, and Joel Mathew. Snow Mountain: Dataset of Audio Recordings of The Bible in Low Resource Languages, May 2023. URL http://arxiv.org/abs/2206.01205. arXiv:2206.01205 [cs, eess].

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.

Revanth Rameshkumar and Peter Bailey. Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5121–5134, Online, July 2020. Association for Computational Linguistics. doi: 10/gtsqxp. URL https://aclanthology.org/2020.acl-main.459.

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.

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A Local-to-Global Approach to Multi-modal Movie Scene Segmentation, April 2020. URL http://arxiv.org/abs/2004.02678. arXiv:2004.02678 [cs].

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset, August 2019. URL http://arxiv.org/abs/1811.00207. arXiv:1811.00207 [cs].

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset, January 2020. URL http://arxiv.org/abs/1909.05855. arXiv:1909.05855 [cs].

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A Conversational Question Answering Challenge, March 2019. URL http://arxiv.org/abs/1808.07042. arXiv:1808.07042 [cs].

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.

Omid Rohanian, Mohammadmahdi Nouriborji, and David A. Clifton. Exploring the Effectiveness of Instruction Tuning in Biomedical Language Processing, December 2023. URL http://arxiv.org/abs/2401.00579. arXiv:2401.00579 [cs].

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A Dataset for Movie Description, January 2015. URL http://arxiv.org/abs/1501.02530. arXiv:1501.02530 [cs].

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description, May 2016a. URL http://arxiv.org/abs/1605.03705. arXiv:1605.03705 [cs].

Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing Fine-Grained and Composite Activities using Hand-Centric Features and Script Data. *International Journal of Computer Vision*, 119(3):346–373, September 2016b. ISSN 0920-5691, 1573-1405. doi: 10/f8w6kp. URL http://arxiv.org/abs/1502.06648. arXiv:1502.06648 [cs].

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.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection, May 2019. URL http://arxiv.org/abs/1901.01342. arXiv:1901.01342 [cs, eess].

Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. XGAN: Unsupervised Image-to-Image Translation for Many-to-Many Mappings, July 2018. URL http://arxiv.org/abs/1711.05139. arXiv:1711.05139 [cs].

Askar Rozi, Dong Wang, Zhiyong Zhang, and Thomas Fang Zheng. An open/free database and Benchmark for Uyghur speaker recognition. In *2015 International Conference Oriental COCOSDA*

held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), pp. 81–85, Shanghai, China, October 2015. IEEE. ISBN 978-1-4673-8279-3. doi: 10/grh5rd. URL http://ieeexplore.ieee.org/document/7357869/.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization, September 2015. URL http://arxiv.org/abs/1509.00685. arXiv:1509.00685 [cs].

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. ISSN 1573-1405. doi: 10/gcgk7w. URL https://doi.org/10.1007/s11263-015-0816-y.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language Modelling with Pixels, April 2023. URL http://arxiv.org/abs/2207.06991. arXiv:2207.06991 [cs].

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of Natural Language Rules in Conversational Machine Reading, August 2018. URL http://arxiv.org/abs/1809.01494. arXiv:1809.01494 [cs, stat].

Matthew J. Sag. The new legal landscape for text mining and machine learning. In *Journal of the Copyright Society of the USA*, 2020.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension, October 2018. URL http://arxiv.org/abs/1804.07927. arXiv:1804.07927 [cs].

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale, November 2019. URL http://arxiv.org/abs/1907.10641. arXiv:1907.10641 [cs].

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The Multilingual TEDx Corpus for Speech Recognition and Translation, June 2021. URL http://arxiv.org/abs/2102.01757. arXiv:2102.01757 [cs].

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.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A Large-scale Dataset for Multimodal Language Understanding, December 2018. URL http://arxiv.org/abs/1811.00347. arXiv:1811.00347 [cs].

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR, March 2023. URL http://arxiv.org/abs/2303.18110. arXiv:2303.18110 [cs, eess].

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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen,

Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask Prompted Training Enables Zero-Shot Task Generalization, March 2022. URL http://arxiv.org/abs/2110.08207. arXiv:2110.08207 [cs].

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. SocialIQA: Commonsense Reasoning about Social Interactions, September 2019. URL http://arxiv.org/abs/1904.09728. arXiv:1904.09728 [cs].

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. Question-Driven Summarization of Answers to Consumer Health Questions, May 2020. URL http://arxiv.org/abs/2005.09067. arXiv:2005.09067 [cs].

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. ARB: Advanced Reasoning Benchmark for Large Language Models, July 2023. URL http://arxiv.org/abs/2307.13692. arXiv:2307.13692 [cs].

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.

Claudia Schulz, Josh Levy-Kramer, Camille Van Assel, Miklos Kepes, and Nils Hammerla. Biomedical Concept Relatedness – A large EHR-based benchmark, October 2020. URL http://arxiv.org/abs/2010.16218. arXiv:2010.16218 [cs].

Abigail See, Peter J. Liu, and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks, April 2017. URL http://arxiv.org/abs/1704.04368. arXiv:1704.04368 [cs].

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a Conversational Agent Overnight with Dialogue Self-Play, January 2018. URL http://arxiv.org/abs/1801.04871. arXiv:1801.04871 [cs].

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis, April 2016. URL http://arxiv.org/abs/1604.02808. arXiv:1604.02808 [cs].

Igor Shalyminov, Sungjin Lee, Arash Eshghi, and Oliver Lemon. Few-Shot Dialogue Generation Without Annotated Data: A Transfer Learning Approach, August 2019. URL http://arxiv.org/abs/1908.05854. arXiv:1908.05854 [cs].

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding Human Hands in Contact at Internet Scale, June 2020. URL http://arxiv.org/abs/2006.06669. arXiv:2006.06669 [cs].

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 664–674, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10/gm8jvb. URL https://aclanthology.org/P18-1062.

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.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding, April 2020. URL http://arxiv.org/abs/2004.06704. arXiv:2004.06704 [cs].

Aidean Sharghi, Jacob S. Laurel, and Boqing Gong. Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach, July 2017. URL http://arxiv.org/abs/1707.04960. arXiv:1707.04960 [cs].

Eva Sharma, Chen Li, and Lu Wang. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization, June 2019. URL http://arxiv.org/abs/1906.03741. arXiv:1906.03741 [cs].

Vivek Sharma, Makarand Tapaswi, and Rainer Stiefelhagen. Deep Multimodal Feature Encoding for Video Ordering, April 2020. URL http://arxiv.org/abs/2004.02205. arXiv:2004.02205 [cs].

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. Highland Puebla Nahuatl Speech Translation Corpus for Endangered Language Documentation. In Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann (eds.), *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 53–63, Online, June 2021. Association for Computational Linguistics. doi: 10/gtwwcd. URL https://aclanthology.org/2021.americasnlp-1.7.

Ying Shi, Askar Hamdullah, Zhiyuan Tang, Dong Wang, and Thomas Fang Zheng. A free Kazakh speech database and a speech recognition baseline. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 745–748, Kuala Lumpur, December 2017. IEEE. ISBN 978-1-5386-1542-3. doi: 10/gtsqzf. URL http://ieeexplore.ieee.org/document/8282133/.

E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41 (Pt 3-4):443–492, 1998. ISSN 0023-8309. doi: 10.1177/002383099804100410.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning, March 2021. URL http://arxiv.org/abs/2010.03768. arXiv:2010.03768 [cs].

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

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, July 2016b. URL http://arxiv.org/abs/1604.01753. arXiv:1604.01753 [cs].

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and Observer: Joint Modeling of First and Third-Person Videos, April 2018. URL http://arxiv.org/abs/1804.09627. arXiv:1804.09627 [cs].

Damien Sileo and Marie-Francine Moens. Probing neural language models for understanding of words of estimative probability, June 2023. URL http://arxiv.org/abs/2211.03358. arXiv:2211.03358 [cs].

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, 2024a.

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, February 2024b. URL http://arxiv.org/abs/2402.06619. arXiv:2402.06619 [cs].

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

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A Short Note on the Kinetics-700-2020 Human Action Dataset, October 2020. URL http://arxiv.org/abs/2010.10864. arXiv:2010.10864 [cs].

Imdat Solak. The M-AILABS Speech Dataset – caito, January 2024. URL https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/.

Per Erik Solberg and Pablo Ortiz. The Norwegian Parliamentary 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, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1003–1008, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.106.

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.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. MAD: A Scalable Dataset for Language Grounding in Videos from Movie Audio Descriptions, March 2022. URL http://arxiv.org/abs/2112.00431. arXiv:2112.00431 [cs].

Amir Soleimani, Christof Monz, and Marcel Worring. NLQuAD: A Non-Factoid Long Question Answering Data Set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1245–1255, Online, 2021. Association for Computational Linguistics. doi: 10/g6k6th. URL https://aclanthology.org/2021.eacl-main.106.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, December 2012. URL http://arxiv.org/abs/1212.0402. arXiv:1212.0402 [cs].

Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, UbiComp '13, pp. 729–738, New York, NY, USA, September 2013. Association for Computing Machinery. ISBN 978-1-4503-1770-2. doi: 10/gtwv9t. URL https://doi.org/10.1145/2493432.2493482.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373, September 2000. ISSN 0891-2017, 1530-9312. doi: 10/dqmv4j. URL http://arxiv.org/abs/cs/0006023. arXiv:cs/0006023.

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.

Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. Conifer: Improving Complex Constrained Instruction-Following Ability of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.02823. arXiv:2404.02823 [cs].

Jennifer J. Sun, Ting Liu, Alan S. Cowen, Florian Schroff, Hartwig Adam, and Gautam Prasad. EEV: A Large-Scale Dataset for Studying Evoked Expressions from Video, February 2021. URL http://arxiv.org/abs/2001.05488. arXiv:2001.05488 [cs].

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. QuaRel: A Dataset and Models for Answering Questions about Qualitative Relationships, November 2018. URL http://arxiv.org/abs/1811.08048. arXiv:1811.08048 [cs].

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. QuaRTz: An Open-Domain Dataset of Qualitative Relationship Questions, September 2019. URL http://arxiv.org/abs/1909.03553. arXiv:1909.03553 [cs].

Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe. JTube-Speech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification, December 2021. URL http://arxiv.org/abs/2112.09323. arXiv:2112.09323 [cs, eess].

Alon Talmor and Jonathan Berant. The Web as a Knowledge-Base for Answering Complex Questions. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10/gkz2k6. URL https://aclanthology.org/N18-1059.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge, March 2019. URL http://arxiv.org/abs/1811.00937. arXiv:1811.00937 [cs].

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex Question Answering over Text, Tables and Images, April 2021. URL http://arxiv.org/abs/2104.06039. arXiv:2104.06039 [cs].

Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. WIQA: A dataset for "What if..." reasoning over procedural text, September 2019. URL http://arxiv.org/abs/1909.04739. arXiv:1909.04739 [cs].

Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. Rapidly Bootstrapping a Question Answering Dataset for COVID-19, April 2020. URL http://arxiv.org/abs/2004.11339. arXiv:2004.11339 [cs].

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, March 2019. URL http://arxiv.org/abs/1903.02874. arXiv:1903.02874 [cs].

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/0336dcbab05b9d5ad24f4333c7658a0e-Abstract-round2.html.

Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. StoryGraphs: Visualizing Character Interactions as a Timeline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 827–834, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Tapaswi_StoryGraphs_Visualizing_Character_2014_CVPR_paper.html.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering, September 2016. URL http://arxiv.org/abs/1512.02902. arXiv:1512.02902 [cs].

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.

Avijit Thawani, Jay Pujara, and Filip Ilievski. Numeracy enhances the Literacy of Language Models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6960–6967, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10/g6k6w5. URL https://aclanthology.org/2021.emnlp-main.557.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for Fact Extraction and VERification, December 2018. URL http://arxiv.org/abs/1803.05355. arXiv:1803.05355 [cs].

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. OpenMathInstruct-1: A 1.8 Million Math Instruction Tuning Dataset, February 2024. URL http://arxiv.org/abs/2402.10176. arXiv:2402.10176 [cs].

Radim Tyleček and Radim Šára. Spatial Pattern Templates for Recognition of Objects with Regular Structure. In Joachim Weickert, Matthias Hein, and Bernt Schiele (eds.), *Pattern Recognition*, pp. 364–374, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-40602-7. doi: 10/ggwb5g.

Herbert Ullrich, Jan Drchal, Martin Rýpar, Hana Vincourová, and Václav Moravec. CsFEVER and CTKFacts: Acquiring Czech data for fact verification. *Language Resources and Evaluation*, 57 (4):1571–1605, December 2023. ISSN 1574-020X, 1574-0218. doi: 10/g6k95h. URL http://arxiv.org/abs/2201.11115. arXiv:2201.11115 [cs].

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. Generative Language Models for Paragraph-Level Question Generation, January 2023. URL http://arxiv.org/abs/2210.03992. arXiv:2210.03992 [cs].

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.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards Understanding Human-Centric Situations from Videos, April 2018. URL http://arxiv.org/abs/1712.06761. arXiv:1712.06761 [cs].

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. Finnish Parliament ASR corpus - Analysis, benchmarks and statistics, March 2022. URL http://arxiv.org/abs/2203.14876. arXiv:2203.14876 [cs, eess].

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, February 2019. URL http://arxiv.org/abs/1804.07461. arXiv:1804.07461 [cs].

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models, January 2022a. URL http://arxiv.org/abs/2111.02840. arXiv:2111.02840 [cs].

Changhan Wang, Anne Wu, and Juan Pino. CoVoST 2 and Massively Multilingual Speech-to-Text Translation, October 2020a. URL http://arxiv.org/abs/2007.10310. arXiv:2007.10310 [cs, eess].

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation, July 2021. URL http://arxiv.org/abs/2101.00390. arXiv:2101.00390 [cs, eess].

Dong Wang and Xuewei Zhang. THCHS-30 : A Free Chinese Speech Corpus, December 2015. URL http://arxiv.org/abs/1512.01882. arXiv:1512.01882 [cs].

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 Open Research Dataset, July 2020b. URL http://arxiv.org/abs/2004.10706. arXiv:2004.10706 [cs].

Wenhan Wang, Ge Li, Bo Ma, Xin Xia, and Zhi Jin. Detecting Code Clones with Graph Neural Networkand Flow-Augmented Abstract Syntax Tree, February 2020c. URL http://arxiv.org/abs/2002.08653. arXiv:2002.08653 [cs].

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

Wenhao Wang and Yi Yang. VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models, September 2024b. URL http://arxiv.org/abs/2403.06098. arXiv:2403.06098 [cs].

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models, February 2024. URL http://arxiv.org/abs/2307.10635. arXiv:2307.10635 [cs].

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research, June 2020d. URL http://arxiv.org/abs/1904.03493. arXiv:1904.03493 [cs].

Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos, March 2022b. URL http://arxiv.org/abs/2203.09463. arXiv:2203.09463 [cs].

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions, May 2023a. URL http://arxiv.org/abs/2212.10560. arXiv:2212.10560 [cs].

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM, November 2023b. URL http://arxiv.org/abs/2311.09528. arXiv:2311.09528 [cs].

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The Benchmark of Linguistic Minimal Pairs for English, February 2023. URL http://arxiv.org/abs/1912.00582. arXiv:1912.00582 [cs].

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.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. AirDialogue: An Environment for Goal-Oriented Dialogue Research. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3844–3854, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10/gf6gq2. URL https://aclanthology.org/D18-1419.

Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards Understanding Human Actions Out of Context, February 2021. URL http://arxiv.org/abs/1912.07249. arXiv:1912.07249 [cs].

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing Multiple Choice Science Questions, July 2017. URL http://arxiv.org/abs/1707.06209. arXiv:1707.06209 [cs, stat].

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing Datasets for Multi-hop Reading Comprehension Across Documents, June 2018. URL http://arxiv.org/abs/1710.06481. arXiv:1710.06481 [cs].

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.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. Learning from Task Descriptions, November 2020. URL http://arxiv.org/abs/2011.08115. arXiv:2011.08115 [cs].

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks, December 2015. URL http://arxiv.org/abs/1502.05698. arXiv:1502.05698 [cs, stat].

Adina Williams, Tristan Thrush, and Douwe Kiela. ANLIzing the Adversarial Natural Language Inference Dataset, October 2020. URL http://arxiv.org/abs/2010.12729. arXiv:2010.12729 [cs].

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

Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1282–1289, Kyoto, Japan, September 2009. IEEE. ISBN 978-1-4244-4442-7. doi: 10/fv4tmg. URL http://ieeexplore.ieee.org/document/5457461/.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA: Towards Building Open-source Language Models for Medicine, August 2023. URL http://arxiv.org/abs/2304.14454. arXiv:2304.14454 [cs].

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A Benchmark for Molecular Machine Learning, October 2018. URL http://arxiv.org/abs/1703.00564. arXiv:1703.00564 [physics, stat].

Wangmeng Xiang, Chao Li, Ke Li, Biao Wang, Xian-Sheng Hua, and Lei Zhang. CDAD: A Common Daily Action Dataset with Collected Hard Negative Samples. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3920–3929, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66548-739-9. doi: 10/gtsqzp. URL https://ieeexplore.ieee.org/document/9857185/.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. TWEETQA: A Social Media Focused Question Answering Dataset, July 2019a. URL http://arxiv.org/abs/1907.06292. arXiv:1907.06292 [cs].

Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A Graph-Based Framework to Bridge Movies and Synopses, October 2019b. URL http://arxiv.org/abs/1910.11009. arXiv:1910.11009 [cs].

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.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. KIWI: A Dataset of Knowledge-Intensive Writing Instructions for Answering Research Questions, March 2024a. URL http://arxiv.org/abs/2403.03866. arXiv:2403.03866 [cs].

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, June 2016. doi: 10/ggv9gj. URL https://ieeexplore.ieee.org/document/7780940. ISSN: 1063-6919.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing, June 2024b. URL http://arxiv.org/abs/2406.08464. arXiv:2406.08464 [cs].

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions, July 2022. URL http://arxiv.org/abs/2111.10337. arXiv:2111.10337 [cs].

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5179–5187, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10/gfsj74. URL http://ieeexplore.ieee.org/document/7299154/.

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.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational(RAMC) Speech Dataset, March 2022. URL http://arxiv.org/abs/2203.16844. arXiv:2203.16844 [cs, eess].

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering, September 2018. URL http://arxiv.org/abs/1809.09600. arXiv:1809.09600 [cs].

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents, February 2023. URL http://arxiv.org/abs/2207.01206. arXiv:2207.01206 [cs].

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. SelFee: Iterative Self-Revising LLM Empowered by Self-Feedback Generation, May 2023. URL https://kaistai.github.io/SelFee/.

Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos, June 2017. URL http://arxiv.org/abs/1507.05738. arXiv:1507.05738 [cs].

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10/gkz3hq. URL https://aclanthology.org/P16-2033.

Yue Yin, Daijiro Mori, and Seiji Fujimoto. ReazonSpeech: A Free and Massive Corpus for Japanese ASR, 2023. URL https://research.reazon.jp/_static/reazonspeech_nlp2023.pdf.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models, October 2023. URL http://arxiv.org/abs/2309.12284. arXiv:2309.12284 [cs].

Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter S. Lasecki, and Dragomir Radev. CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases, September 2019a. URL http://arxiv.org/abs/1909.05378. arXiv:1909.05378 [cs].

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task, February 2019b. URL http://arxiv.org/abs/1809.08887. arXiv:1809.08887 [cs].

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. SParC: Cross-Domain Semantic Parsing in Context. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4511–4523, Florence, Italy, July 2019c. Association for Computational Linguistics. doi: 10/gj4fwd. URL https://aclanthology.org/P19-1443.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning, August 2020. URL http://arxiv.org/abs/2002.04326. arXiv:2002.04326 [cs].

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning, October 2023. URL http://arxiv.org/abs/2309.05653. arXiv:2309.05653 [cs].

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines, July 2020. URL http://arxiv.org/abs/2007.12720. arXiv:2007.12720 [cs].

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence?, May 2019. URL http://arxiv.org/abs/1905.07830. arXiv:1905.07830 [cs].

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling Generalized Agent Abilities for LLMs, October 2023. URL http://arxiv.org/abs/2310.12823. arXiv:2310.12823 [cs].

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Title Generation for User Generated Videos, September 2016. URL http://arxiv.org/abs/1608.07068. arXiv:1608.07068 [cs].

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition, February 2022a. URL http://arxiv.org/abs/2110.03370. arXiv:2110.03370 [cs].

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 2023a.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. Chinese Open Instruction Generalist: A Preliminary Release, April 2023b. URL http://arxiv.org/abs/2304.07987. arXiv:2304.07987 [cs].

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and Philip S. Yu. Are Pretrained Transformers Robust in Intent Classification? A Missing Ingredient in Evaluation of Out-of-Scope Intent Detection, April 2022b. URL http://arxiv.org/abs/2106.04564. arXiv:2106.04564 [cs].

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence, March 2023c. URL http://arxiv.org/abs/2209.02970. arXiv:2209.02970 [cs].

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification, April 2016. URL http://arxiv.org/abs/1509.01626. arXiv:1509.01626 [cs].

Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry Davis. VideoLT: Large-scale Long-tailed Video Recognition, August 2021. URL http://arxiv.org/abs/2105.02668. arXiv:2105.02668 [cs].

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. AlpaCare:Instruction-tuned Large Language Models for Medical Application, April 2024. URL http://arxiv.org/abs/2310.14558. arXiv:2310.14558 [cs].

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 2022c. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3182537. URL http://dx.doi.org/10.1109/JSTSP.2022.3182537.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling, April 2019. URL http://arxiv.org/abs/1904.01130. arXiv:1904.01130 [cs].

Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization, September 2019. URL http://arxiv.org/abs/1712.09374. arXiv:1712.09374 [cs].

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *Proceedings of the Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 868–884, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4. doi: 10/gtwv9w.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. URL http://arxiv.org/abs/2306.05685. arXiv:2306.05685 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset, March 2024a. URL http://arxiv.org/abs/2309.11998. arXiv:2309.11998 [cs].

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 2024b. URL https://github.com/hpcaitech/Open-Sora.

Tiankai Zhi, Ying Shi, Wenqiang Du, Guanyu Li, and Dong Wang. M2ASR-MONGO: A Free Mongolian Speech Database and Accompanied Baselines. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 140–145, Singapore, Singapore, November 2021. IEEE. ISBN 978-1-66540-870-7. doi: 10/gtsqzg. URL https://ieeexplore.ieee.org/document/9660401/.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization, April 2021. URL http://arxiv.org/abs/2104.05938. arXiv:2104.05938 [cs].

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning, November 2017. URL http://arxiv.org/abs/1709.00103. arXiv:1709.00103 [cs].

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards Automatic Learning of Procedures from Web Instructional Videos, March 2017. URL https://arxiv.org/abs/1703.09788v3. arXiv:1703.09788 [cs].

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. DocPrompting: Generating Code by Retrieving the Docs, February 2023a. URL http://arxiv.org/abs/2207.05987. arXiv:2207.05987 [cs].

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements, June 2023b. URL http://arxiv.org/abs/2306.01985. arXiv:2306.01985 [cs].

Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, and Yang Liu. Devign: Effective Vulnerability Identification by Learning Comprehensive Program Semantics via Graph Neural Networks, September 2019. URL http://arxiv.org/abs/1909.03496. arXiv:1909.03496 [cs, stat].

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization, March 2021. URL http://arxiv.org/abs/2103.06410. arXiv:2103.06410 [cs].

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.

Ming Zhu, Aneesh Jain, Karthik Suresh, Roshan Ravindran, Sindhu Tipirneni, and Chandan K. Reddy. XLCoST: A Benchmark Dataset for Cross-lingual Code Intelligence, June 2022. URL http://arxiv.org/abs/2206.08474. arXiv:2206.08474 [cs].

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.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos, April 2019. URL http://arxiv.org/abs/1903.08225. arXiv:1903.08225 [cs].