The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources

Anonymous authors
Paper under double-blind review

Abstract

Foundation model development attracts a rapidly expanding body of contributors, scientists, and applications. To help shape *responsible development practices*, we introduce the Foundation Model Development Cheatsheet: a growing collection of 250+ tools and resources spanning text, vision, and speech modalities. We draw on a large body of prior work to survey resources (*e.g.* software, documentation, frameworks, guides, and practical tools) that support informed data selection, processing, and understanding, precise and limitation-aware artifact documentation, efficient model training, advance awareness of the environmental impact from training, careful model evaluation of capabilities, risks, and claims, as well as responsible model release, licensing and deployment practices. We hope this curated collection of resources helps guide more responsible development. The process of curating this list, enabled us to review the AI development ecosystem, revealing what tools are critically missing, misused, or over-used in existing practices. We find that (i) tools for data sourcing, model evaluation, and monitoring are critically under-serving ethical and real-world needs, (ii) evaluations for model safety, capabilities, and environmental impact all lack reproducibility and transparency, (iii) text and particularly English-centric analyses continue to dominate over multilingual and multi-modal analyses, and (iv) evaluation of systems, rather than just models, is needed so that capabilities and impact are assessed in context.

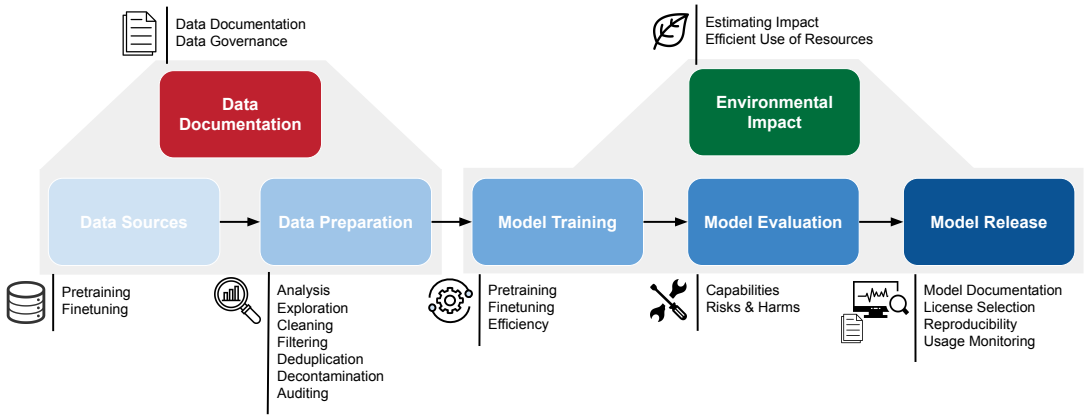

Figure 1: **The phases of model development by which this survey and review paper is organized.**

1 Introduction

As the capabilities (Üstün et al., 2024; Team et al., 2023; Gomez, 2024; Anthropic, 2024a; Radford et al., 2023; Brooks et al., 2024) and market prospects (Vipra & Korinek, 2023; McElheran et al., 2024) of artificial intelligence have quickly expanded, so have the communities of developers, scientists, and contributors who build foundation models (Bommasani et al., 2021). The fields' growth has spurred widespread adoption of many tools and resources used to build, deploy, evaluate, and govern large foundation models (FMs). However, these nascent practices are often immature. Many outstanding resources are neglected, in part for lack of discoverability, or awareness of good practices. To address these gaps, we conduct a focused survey, not of scientific literature (which already exists for many FM development topics (Albalak et al., 2024; Zhao et al., 2023a; Chang et al., 2023)), but of *resources* for

FM development such as datasets, software packages, documentation, guides, frameworks, and practical tools. In particular, this resource curation is tailored to responsible development practices for newer, smaller, or mid-sized development teams. Large FM development organizations, such as Google, OpenAI, or Meta, with substantial user bases, should adhere to more rigorous and product-specific best practices than outlined in this review. We release the Foundation Model Development Cheatsheet, the repository of annotated tools for text, speech, and vision models, and open it for public contributions.

For each phase of model development, our contributions are summarized as (i) a survey of relevant tools and resources, (ii) a synthesis of the literature's recommended practices and use of those tools, and (iii) a review of the limitations and omissions of existing resources. The cheatsheet serves as a succinct guide of the survey and recommended practices, prepared *by* foundation model developers *for* foundation model developers. The intended audience is a range of foundation model developers, including academic researchers, startup companies, research labs, who are pretraining from scratch, or simply finetuning, big and small.

Our survey and recommendations hope to bring wide attention to tools across several phases of development. First, we suggest resources that support *informed* data selection, processing, and understanding (§§ 3 and 4). Data prepared without sufficient due diligence can lead to unintended consequences (e.g. risks to privacy, copyright, or generating sensitive content), marginalization (e.g. by inadvertently filtering out certain distributions), or unexpected model behaviors (e.g. train/test overlap or security vulnerabilities). Next we survey resources for *precise and limitation-aware* artifact documentation (Section 5). When new datasets are released, setting their governance standards early will avoid misuse later. Foundation model training can be financially and environmentally expensive. We aggregate resources for *efficient* model training (Section 6) and estimating a model's scaling behavior and environmental impact (Section 7). *Advance awareness* of these quantities can inform more efficient training practices. For once models are trained, we provide evaluation frameworks, taxonomies of risk, and benchmarks for a variety of evaluation criteria (Section 8). Best practices suggest models should be evaluated for their intended uses, as well as some unforeseen misuses or harms. Developers should design naturalistic evaluations for these settings rather than relying on available but poorly fitting tools (Biderman et al., 2024b; Liao & Xiao, 2023). And our evaluation frameworks suggest evaluation metrics should be contextualized, to avoid over-claiming or misunderstanding the limitations of the reported numbers. Lastly, our survey informs *responsible* model release and deployment practices (Section 9), so developers can make informed selections of licenses and release mechanisms, to address misuse risks.

Beyond the survey of resources, we review the existing ecosystem of tools and resources. For each segment of model development, we examine the limitations, omissions, and opportunities for improvement of existing tooling and common practices. Summarized in Table 1, we find:

- Tools for data sourcing, model evaluation, and monitoring are critically under-serving responsible development needs and real-world needs. For instance, they often fail to have sufficient documentation, imitate real use cases, or accurately reflect licensing permissions.
- Popular model safety, capabilities, and environmental impact evaluation benchmarks lack reproducibility and transparency.
- Resources for multilingual and multi-modal development, across every phase of development, continue to receive significantly less attention than English and text-centric equivalents.
- Resources to evaluate *systems*, rather than just models, is needed so that capabilities and impact are assessed in the context of real-world deployment.

Development Phase	Frequent Status Quo	Review & Recommendations
All Phases	Predominantly English, and text-centric resources.	More multilingual, multi-modal, and flexible resources.
	A concentration of standalone, flashy, one-off projects, for credit incentives.	More collaborative, large-scale interoperable infrastructure projects that builds on existing tooling.
Data Sources	Predominantly synthetic, unrealistic finetuning data.	Naturalistic observations and realistic training tasks.
	Predominantly English, and text-centric data sources.	More multilingual, multi-modal data sources.
	Intended uses, licenses, consent, & provenance are haphazardly documented.	Prioritizing datasets with standardized, structured, and linked metadata.
	Sparse information especially on data's sensitive content, such as CSAM and NCII.	Comprehensive "data measurements" as part of large, unstructured data releases.
	Data & modeling focus on easy-to-scale formats, e.g. captioned images.	More attention to collecting neglected data formats, e.g. naturally interspersed text and images.
Data Preparation	Diverse data infrastructure and processing standards for individual use-cases.	Interoperable data formats and processing for many use-cases.
	Mostly one-off and closed-source data exploration tools.	More open data exploration tools.
	Loose ideas of "high-quality" data that applies to all domains.	Concrete and precise definitions of quality that are unique to a set of domains, tasks, or evaluations.
	Data preparation methods compared across models with different training data.	Standardized data-centric benchmarks to fairly compare methods.
	English-centric tokenization and processing tools.	More multilingual and low-resource language tokenization and processing tools.
Data Documentation	Data documentation is diversely formatted, often terse, performative, without achieving reproducibility.	Executable, and verifially reproducible scraping and analysis scripts. Standardized data documentation requirements (e.g. from conferences).
	Documentation is an after-thought.	Documentation is started early, assembled over the course of collection and processing.
	Data stewardship and maintenance are often ignored beyond initial release.	Data governance is proactively organized, with a post-launch maintenance and licensing plan.
Model Training	Mostly educational resources for technical developers.	More "last mile" educational resources for non-technical developers.
	Inflexible and disparate tooling.	Standardized and centralized resources, especially cross-modality.
Environmental Impact	Opaque environmental impact estimate programs from cloud providers.	Query-level energy and environmental usage transparency from .
	Opaque information from hardware makers and data centers.	Fine-grained transparency from hardware makers and data centers.
	Inability to compare energy standards across systems.	"Energy Star" standards for non-technical users to fairly compare AI services.
	Scaling laws are text-centric and impractical.	Multimodal scaling laws research. Intuitive interfaces to estimate and model scaling laws for training forecasts.
Model Evaluation	Evaluating model outputs.	Evaluating systems, within their application contexts.
	Evaluating on synthetic toxicity and safety benchmarks.	Evaluating natural observations, real-world settings, and with human-interaction studies.
	Reporting evaluation metrics only.	Releasing evaluation scripts for verifiable reproducibility.
Model Release	No or uninformed license choices.	License selection guided by the context of data, potential and unforeseen uses, and legal considerations.
	Weights release with limited support.	Accompanying documentation, and easy-to-run code for training, evaluation, and inference.
	Limited plans for usage monitoring, or over-claimed benefits from watermarking/monitoring.	A plan that considers gating, watermarking, and misuse reporting (though they are not always beneficial).
	Harm & hazard taxonomies are based on existing benchmarks.	Harm taxonomies are based on empirical observations, and studies with real users.

Table 1: **A summary of the reviews & take-aways from each phase of foundation model development ecosystem.**

2 Methodology & Guidelines

We develop the following methodology to guide the collection of these tools and resources, as well as our analysis. First, we've divided the model development pipeline into several phases, illustrated in Figure 1. Despite the visual representation, we acknowledge these phases are frequently interrelated rather than sequential. We also include categories that have been identified by existing scholarship as necessary to responsible development, even though they are frequently omitted from development pipelines: such as documentation, environmental impact estimation, or risks and harms evaluation. Next, authors are allocated to these phases based on their areas of expertise, or to modalities (text, vision, or speech) across a few phases of development. Each author surveys the literature in their segment of model development to find a mix of (a) relevant tools, in the form of repositories, APIs, or interfaces, (b) scientific literature that directly surveys or guides development decisions, and (c) frameworks or standards for development (such as documentation or risk taxonomies). Due to the breadth of categories, the literature survey is inevitably non-exhaustive, but reflects a dedicated search for the most prominent tools in each area. Certain phases of development, such as model training, have thousands of available repositories, so we curate a criteria for inclusion (outlined below) that prioritizes certain qualities: popularity, usefulness, and advancing responsible practices.

Criteria for Inclusion. These principles for inclusion, described below, are incomplete and subjective, but we still believe sufficiently rigorous to make for a thorough and useful compilation of resources. The resources are selected based on a literature review for each phase of foundation model development. Inclusion is predicated on a series of considerations, including: the popularity of the tool on Hugging Face or GitHub, the perceived helpfulness as a development tool, the extent and quality of the documentation, the insights brought to the development process, and, in some cases, the lack of awareness a useful resource has received in the AI community. For an example of this last consideration, in Section 3.2 we try to include more Finetuning Data Catalogs for lower resource languages, that often receive less attention than ones featured more prominently on Hugging Face's Dataset Hub. Rather than sharing primarily academic literature as in most surveys, we focus on tools, such as data catalogs, search/analysis tools, evaluation repositories, and, selectively, literature that summarizes, surveys or guides important development decisions. Further, we hope to make the coverage more comprehensive with an open call for community contributions.

Scope & Limitations. We've compiled resources, tools, and papers that guide model development, and which we believe will be especially helpful to nascent (and often experienced) foundation model developers. However, this guide is *far from exhaustive*—and here's what to consider when using it:

- **Temporal scope.** Foundation model development is a rapidly evolving science. **This document is only a sample, dated to March 2024.**
- **Applicable developers.** We scope these resources to mid-to-small foundation model developers. Large organizations, with commercial services and/or wide user bases, have broader considerations for responsible development that what is outlined in this work.
- **Development phases & modalities** We've scoped our data modalities only to **text, vision, and speech**, and to the phases of development outlined in Figure 1. We support multilingual resources, but acknowledge there remain significant community-wide gaps in awareness and adoption here..
- **We remind users to follow standard practice in assessing the security and viability of each tool, particularly for their circumstance.** At times we have provided resources with conflicting advice, as it is helpful to be aware of divided community perspectives. Our notes throughout are designed to contextualize these resources, to help guide the reader's judgement.

3 Data Sources

Data Sourcing Best Practices

- Pretraining data provides the fundamental ingredient to foundation models—including their capabilities and flaws. Finetuning data improves the model's performance in specific settings, or in the case of instruction finetuning or alignment training, improves the model's general usability and helpfulness while aiming to reduce potential harms.
- More data is not always better. It is essential to carefully source data, and manually inspect it to ensure it *fits the goals of your project.*
- Dataset selection includes many relevant considerations, such as language and dialect coverage, topics, tasks, diversity, quality, and representation.
- Most datasets come with implicit modifications and augmentations, from their selection, filtering, and formatting. Pay attention to these pre-processing steps, as they will impact your model.
- Finetuning data can improve the model's performance in some settings or impair others. Use catalogs to support an informed selection, and prefer well-documented to under-documented datasets.
- Crowdsourced data catalogs, including HuggingFace, may contain important omissions and errors in their dataset documentation. Verify information, such as data licensing and critical characteristics, with the original data sources and academic papers, if possible.
- The most appropriate datasets may not exist for a given set of tasks. Be aware of the limitations of choosing from what is available.

3.1 Pretraining Data Sources

Pretraining corpora consist of millions of pieces of content, from documents, images, videos, or speech recordings, often scraped directly from the web. Model pretraining represents the fundamental step in instilling foundation models with their abilities to represent syntax, grammar, reasoning, and world knowledge (Devlin et al., 2018; Brown et al., 2020; Chowdhery et al., 2023). Consequently, it is important to carefully curate the data composition, including the mix of sources, characteristics, and preprocessing decisions (Longpre et al., 2023b). However, the vast scale of this content often means its is shallowly documented and understood, despite community efforts to unpack it (Dodge et al., 2021; Elazar et al., 2023a).

We highlight a few of the most popular pretraining corpora which have accumulated deeper documentation and analysis. In the text domain, web scrapes from common crawl (`commoncrawl.org`), or OSCAR (`https://oscar-project.org/`) (Suárez et al., 2019; Laippala et al., 2022) are the base ingredient for most pretraining corpora. In particular, derivatives of common crawl, such as C4 (Raffel et al., 2020; Dodge et al., 2021) or multilingual C4 (Kreutzer et al., 2022) provide 2019 indexes that have been heuristically filtered for well-formed text. Subsequent pretraining datasets incorporate one or multiple indexes from common crawl. This includes the Pile (Gao et al., 2020), RefinedWeb (Penedo et al., 2023), RedPajama (Together AI, 2023), and Dolma (Soldaini et al., 2024), which have iterated on basic document quality filtering and deduplication. These corpora often merge and deduplicate multiple years of common crawl scrapes, or supplement additional sources including biomedical text (PubMed), legal text (Freelaw, USPTO patent documents), code (Github, Stack Exchange), public domain books (Project Gutenberg), among others.

For multilingual text, the Open Parallel Corpus(OPUS) offers a massive collection of translated text document pairs (Tiedemann, 2012), ROOTS (Laurençon et al., 2022) collates and processes diverse multilingual resources, including OSCAR (Laippala et al., 2022) and the Bigscience Catalogue (McMillan-Major et al., 2022), CulturaX (Nguyen et al., 2023) covers 167 languages from OSCAR and mC4, and WURA (Oladipo et al., 2023) centralizes and manually audits documents from 16 African languages. Most recently, MADLAD-400 (Kudugunta et al., 2023) provides a 3 trillion token, 2023 processed split of Common Crawl, spanning 419 languages.

Large, specialized corpora of text have also recently emerged, to specialize model abilities, or mitigate risks of possible copyright infringement. As examples, the Pile of Law (Henderson et al., 2022) and MultiLegalPile (Niklaus et al., 2023) centralize court opinions, contracts, and legislative records; the Stack and Stack v2 scrape permissively-

licensed GitHub repositories (Kocetkov et al., 2022); peS2o (Lo et al., 2020) releases cleaned academic papers from Semantic Scholar; and the Proof Pile 2 (Azerbayev et al., 2023) and OpenWebMath (Paster et al., 2023) aggregate vast mathematical text resources. Notably, recent text corpora also *attempt* to isolate permissively-licensed, or copyright free data, such as the Open License Corpus (Min et al., 2023)—though this is a challenging task, and does not guarantee that all non-commercially licensed or copyrighted content has been removed.

In the context of speech, webscrapes of English audiobook data from LibriVox (a website hosting free public domain audiobooks) are commonly used for foundation model architecture development and evaluation. Sourcing data from LibriVox, LibriSpeech (Panayotov et al., 2015) is a 960 hour fully supervised dataset (i.e. all audio are paired with transcriptions) and Libri-Light (Kahn et al., 2020) is a 60k hour dataset for benchmarking using no or limited supervision (10h, 1h, and/or 10min). Multilingual models are typically pre-trained on a combination of speech corpora, e.g. for XLS-R (Babu et al., 2022) a combination 436k hours from VoxPopuli (400k hours from 23 languages based on European Parliament recordings: Wang et al., 2021), Multilingual LibriSpeech (50k hours of audiobooks from 8 languages Pratap et al., 2020), CommonVoice (28k hours of crowd-sourced read speech from 100+ languages Ardila et al., 2020), VoxLingua107 (6.6k hours from 107 languages scraped from YouTube Valk & Alumäe, 2021), and various IARPA BABEL corpora (totaling 1k hours from 17 African and Asian languages).

In the context of vision, there are several large pretraining corpora, comprised of text and images found in large web scrapes. COYO aggregates 700M images with alt-text from the web.[1] Multimodal C4, or MMC4, (Zhu et al., 2024) leverages the original C4 URLs to extract interleaved image and text pairs, totaling 570M images, and 43B tokens. Similarly, OBELICS also leverages the Common Crawl web collection, filters for multimodal web pages, and extracts images from the HTML (Laurençon et al., 2024a), totaling 353M images, with 115B tokens. Lastly, DataComp-1B and CommonPool-13B (Gadre et al., 2024a) isolate high quality subsets of image-text pairs on CommonCrawl. For video specifically, WebVid (Bain et al., 2021) provides 10M videos and their text, from which image datasets can also be extracted. For vision or speech, WebDatasets provides a high-performance data streaming tool.[2]

3.2 Finetuning Data Catalogs

In this section, we survey sources that catalog finetuning datasets—sometimes known more broadly post-training datasets. Finetuning data, differs from pretraining data in that it is curated for supervised learning, is more specialized to the intended inference-time task, and is typically much smaller in scale. Finetuning datasets are used for a variety of reasons: to hone specific capabilities, orient the model to a certain task format, improve its responses to general instructions, mitigate harmful or unhelpful response patterns, or generally align its responses to human preferences. Developers increasingly vary the types of data annotations and loss objectives, depending on the goal. Notably, after pretraining practitioners commonly use traditional supervised finetuning, DPO (Rafailov et al., 2023) or reinforcement learning objectives from human (or machine) feedback to generated responses (Ouyang et al., 2022; Bai et al., 2022).

Given the thousands of specialized data sources for finetuning, we recommend using data catalogs that provide well documented datasets, to make for an informed selection. HuggingFace Datasets (`https://huggingface.co/docs/datasets/index`) offers the largest and most popular AI community catalog across modalities and tasks (Lhoest et al., 2021a). Many datasets are accompanied by data cards, however Longpre et al. (2023b) show there are frequent omissions and errors, as is the nature of crowdsourced catalogs.

There exist an array of other specialized data catalogs with datasets that may not appear directly on HuggingFace. For instance, the NusaCrowd catalog for South East Asian languages (Cahyawijaya et al.), the Masader Arabic data catalogue (Alyafeai et al., 2022), the AI4Barat Indian data catalog (`https://huggingface.co/ai4bharat`), as well as the Maskahane NLP (`https://github.com/masakhane-io`) and Zenodo AfricaNLP catalogs (`https://zenodo.org/communities/africanlp/`) offer specialized multilingual text and speech resources for the language communities. Recently, the Aya Collection (Singh et al., 2024) offers curated, multilingual text finetuning datasets across 65 languages. For more accurate and comprehensive data documentation, particularly for licensing and provenance, the Data Provenance Initiative publishes tools to search and filter popular HuggingFace finetuning datasets (text, speech, and vision) across a variety of criteria (Longpre et al., 2023b). Liu et al. (2024c) recommend best practices in developing and using synthetic data, which has become increasingly popular in the text domain.

[1] `https://huggingface.co/datasets/kakaobrain/coyo-700m`
[2] `https://github.com/webdataset/webdataset`

In the context of speech, OpenSLR (http://openslr.org) is a large collection of user-contributed datasets for various speech processing tasks. For the task of spoken language identification, VoxLingua107 (Valk & Alumäe, 2021) comprises audio scraped from YouTube using various language-specific keywords and, analogously, for speaker identification/verification VoxCeleb (Nagrani et al., 2017) comprises audio of from 1,000 celebrities.

In the context of vision, there are a few well known data sources for finetuning. ImageNet (Deng et al., 2009) provides the historical framework for which image classification models competed on 1.3M samples with 1000 diverse classes. MS COCO (Lin et al., 2014) provides training data for image detection, segmentation, captioning and retrieval. There exist many options for modern instruction finetuning datasets in the vision domain. A few notable examples include the Multi-Modal, Multilingual Instruction Tuning (M3IT) dataset (Li et al., 2023c), comprising 40 datasets, 2.4 million examples over 400 tasks in 80 languages; The Cauldron, comprising 50 vision-language datasets (Laurençon et al., 2024b); and LLaVA Visual Insruct 150k, a set of text-image datasets generated by prompting the GPT-4-0314 API (Liu et al., 2024a). For using web-scraped data, Child and Sexual Abuse Material (CSAM) is an acute concern. Tools such as PhotoDNA can be used to help detect and filter for these images, though they may not be completely accurate or comprehensive.[3]

3.3 Review

In this section we critically review the current state of resources for data sourcing, from our survey.

The community would benefit from more accurate and comprehensive licensing, provenance, and creator consent information for existing datasets. Many datasets tend to be under-documented (Bandy & Vincent, 2021; Sambasivan et al., 2021), or erroneously documented (Longpre et al., 2023b), with 65% of HuggingFace dataset licenses either omitted or incorrectly labelled. This is especially true of large data collections that have re-packaged and sometimes re-licensed hundreds of diverse datasets, each with different documentation standards (Longpre et al., 2023a; Sanh et al., 2021). The tools used to discover, select, and verify the dataset properties are under-developed, especially with respect to concerns of creator consent, copyright infringement, and related terms of use. For creator consent, initial opt-in/opt-out tooling has yet to be widely adopted. While HuggingFace has integrated Spawning's opt-in database (https://api.spawning.ai/spawning-api), there remains limited creator and developer adoption. For copyright information new datasets and catalogs such as the Data Provenance Initiative (Longpre et al., 2023b), offer more detailed license tracing tools, but their coverage also remains limited. Lastly, synthetic data generation (usually using OpenAI APIs) has expanded significantly for finetuning datasets (Longpre et al., 2023b), but don't always document the limitations on the data use imposed by the APIs. This is further compounded by the fact that there is substantial uncertainty about the extent to which the terms of service of such platforms bind downstream users. The absence of infrastructure to trace and verify these types of data documentation lead to uninformed data sourcing and use.

The community would benefit from more accurate and comprehensive information on sensitive content, such as CSAM and NCII. Prior work highlights the risks of proliferated CSAM and NCII as a fundamental risk from generative AI models (Kapoor et al., 2024; Lakatos, 2023b; Thiel et al., 2023a). However, significant portions of the AI community have frequently trained on large-scale datasets that contain this sensitive content without widespread awareness—such as LAION-5B (Birhane et al., 2021; David, 2023). With a greater focus on generative, multimodal, and more memorization-prone large models, there is a greater need for resources to help identify and filter for sensitive content.

Data & modeling have focused predominantly on easy-to-scale data formats, neglecting other formats. A prominent focus of multimodal modeling has been image-to-caption and caption-to-image tasks. The ease of sourcing and scaling these caption datasets to tens of billions has enabled these tasks to progress. However, this has neglected more complex and useful reasoning tasks, that may require text and images with different relationships, interleaved in different ways. In the context of speech, many automatic speech recognition datasets comprise only of read speech (e.g. as collection involved soliciting crowd-sourced participants to read various text stimuli), which is vastly different from informal, multi-speaker conversations which are the "primordial home of human language" (Dingemanse & Liesenfeld, 2022).

There is a scarcity of realistic training tasks, and naturalistic observations. Many open academic datasets are developed for niche, even artificial purposes (e.g. academic visual question datasets that are often detached from real world use cases). To collect realistic use cases, however, requires access to volunteered user logs from real

[3] https://www.microsoft.com/en-us/photodna

products and services. Some attempts have been made to do this, such as WildChat (Zhao et al., 2023b), however their participation may still skew to a superficial distribution. Tools and even policy to scalably source real data, while preserving privacy, is critical to sourcing grounded and relevant training data.

4 Data Preparation

Data Preparation Best Practices

- Tools for **searching and analysing** can help developers better understand their data, and therefore understand how their model will behave; an important, but often overlooked, step of model development.
- Data **cleaning and filtering** can have an immense impact on the model characteristics, though there is not a one size fits all recommendation. The references provide filtering suggestions based on the application and communities the model is intended to serve.
- When training a model on data from multiple sources/domains, the quantity of data seen from each domain (**data mixing**) can have a significant impact on downstream performance. It is common practice to upweight domains of "high-quality" data; data that is known to be written by humans and has likely gone through an editing process such as Wikipedia and books. However, data mixing is an active area of research and best practices are still being developed.
- **Removing duplicated data** can reduce undesirable memorization and can improve training efficiency.
- It is important to carefully **decontaminate training datasets** by removing data from evaluation benchmarks, so their capabilities can be precisely understood.

4.1 Data Search, Analysis, and Exploration

A critical step to understanding a dataset is to explore and analyze what it contains. In particular, exploring training datasets with search and analysis tools can help practitioners develop a nuanced intuition for what exists in the data, and therefore can help to predict what behaviors the model will exhibit.

Tools for search, analysis, and exploration can take a few forms. Some tools are aimed at understanding the high-level statistics of a dataset such as the length of inputs, frequency of specific n-grams, the languages in the corpus, possible biases, or the existence of undesirable content. For example, tools such as WIMBD (Elazar et al., 2023a) and Infini-gram (Liu et al., 2024b) allow users to perform n-gram searches through commonly used pretraining datasets, and provide starting points for building a search index over any arbitrary dataset. The ROOTS search tool [4] (Piktus et al., 2023) additionally allows users to search with fuzzy n-grams over the ROOTS corpus, and similarly, the clip-retrieval tool [5](Beaumont, 2022) allows users to search for nearest neighbor images and text from a multimodal corpus (e.g. LAION-5B (Schuhmann et al., 2022a). Furthermore, the HuggingFace Data measurements tool[6] gives users access to statistics such as the most common n-grams, lengths of data points, and distribution over labels for a number of pretraining and finetuning datasets. Yet more tools exist to explore the data manually and including the Data Provenance Explorer (Longpre et al., 2023b) for text datasets, Google's Know Your Data tool [7] for vision datasets and NVIDIA's Speech Data Explorer [8] for speech datasets.

In most cases, it is highly recommended to explore datasets from the perspective of high-level statistics as well as getting to know the data by looking at individual data points. For instance, text data can have a wide distribution of lengths, topics, tones, formats, licenses, and even diction, and understanding each of these dimensions will require a different tool. We recommend that developers use the many available tools to search and analyze their datasets.

[4] https://huggingface.co/spaces/bigscience-data/roots-search
[5] https://github.com/rom1504/clip-retrieval
[6] https://huggingface.co/spaces/huggingface/data-measurements-tool
[7] https://knowyourdata.withgoogle.com/
[8] https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/tools/speech_data_explorer.html

4.2 Data Cleaning, Filtering, and Mixing

Once the contents of a dataset are understood, the next step is to clean and filter the dataset to adjust the dataset's distribution towards desirable content. Filtering and cleaning do so by removing unwanted data from the dataset. They can improve training efficiency as well as ensure that data has desirable properties, including: high information content, desired languages, low toxicity, and minimal personally identifiable information. Data mixing is another important component of data preparation, where the mixture proportions of data domains (e.g. scientific articles, GitHub, and books) have been shown to dramatically affect downstream performance (Gao et al., 2020; Xie et al., 2023a; Albalak et al., 2023).

A first step for filtering text data is by language, where there is a plethora of tools including langdetect [9], cld3 [10], OpenLID [11] (Burchell et al., 2023), GlotLID [12] (Kargaran et al., 2023), and FastText [13] (Grave et al., 2018). The majority of modern language identification methods have been built on top of the FastText model used in the CCNet pipeline (Wenzek et al., 2020). In addition to filtering by language, datasets are commonly cleaned using heuristics (e.g. remove documents with fewer than 5 words, or remove lines that start with "sign-in") which have been implemented through a number of different tools including the Dolma toolkit [14] (Soldaini et al., 2023), Lilac [15], and DataTrove [16](Penedo et al., 2024), as well as DataComp [17] (Gadre et al., 2023) for image-text pairs. While heuristic filtering methods can remove significant quantities of data, they can be brittle. Model-based methods (e.g. remove data which is dissimilar to a known "high-quality" corpus or possibly toxic content) such as DSIR [18] (Xie et al., 2023b) and Detoxify [19] (Hanu & Unitary, 2020) can be used as additional filtering that allow for much more flexibility than heuristics. In addition to cleaning and filtering, mixing is another component of dataset design which requires careful consideration. Many dataset mixing ratios have been determined by heuristics and human judgement. For example, the Pile (Gao et al., 2020) and Llama (Touvron et al., 2023) upweight domains that have likely gone through an editing process, such as books and Wikipedia articles. Tools for automated data mixing are, as of writing, very limited. However, academic research projects such as DoReMi [20] (Xie et al., 2023a) and Online Data Mixing [21] (Albalak et al., 2023) provide GitHub repositories that can be repurposed for new datasets.

Cleaning, filtering, and mixing are crucial components of designing an appropriate dataset, but due to the vast space of possible filters and downstream uses of ML models there is no one-size-fits-all recommendation. Practitioners looking to clean and filter their datasets should first use the search, analysis, and exploration tools to determine how to design appropriate filters, and iteratively improve the filtering. For more details and recommendations on cleaning, filtering, and mixing, see the recent survey by Albalak et al. (2024).

4.3 Data Deduplication

Data deduplication is an important preprocessing step where duplicated documents, or chunks within a document, are removed from the dataset. Removing duplicates can reduce the likelihood of memorizing undesirable pieces of information such as boilerplate text, copyrighted data, and personally identifiable information. Additionally, removing duplicated data improves training efficiency by reducing the total dataset size.

Deduplication generally relies on one of four methods: URL matching, hashing methods, string metrics, or model representations, which can classify data as either exact or fuzzy matches. We suggest using the deduplication methods included in DataTrove [22](Penedo et al., 2024), Google's deduplicate-text-datasets library [23] (Lee et al., 2022a), or the

[9] https://github.com/Mimino666/langdetect
[10] https://github.com/google/cld3
[11] https://github.com/laurieburchell/open-lid-dataset
[12] https://github.com/cisnlp/GlotLID
[13] https://huggingface.co/facebook/fasttext-language-identification
[14] https://github.com/allenai/dolma
[15] https://github.com/lilacai/lilac
[16] https://github.com/huggingface/datatrove
[17] https://www.datacomp.ai/
[18] https://github.com/p-lambda/dsir
[19] https://github.com/unitaryai/detoxify
[20] https://github.com/sangmichaelxie/doremi
[21] https://github.com/alon-albalak/online-data-mixing
[22] https://github.com/huggingface/datatrove
[23] https://github.com/google-research/deduplicate-text-datasets

Dolma toolkit [24] (Soldaini et al., 2023) for their ease of use. In practice, multiple steps of deduplication are performed. First, simple deduplication can be performed (e.g. URL-based filtering), followed by more complicated hashing- and model-based methods. Practitioners should always determine whether duplicated data will harm or help the model for their use case. While memorization is commonly cast as a bad thing in machine learning, it can be a positive such as when a model "memorizes" the answer to a factual question (Biderman et al., 2024a).

4.4 Data Decontamination

Data decontamination is the process of removing evaluation data from the training dataset. This important step in data preprocessing ensures the integrity of model evaluation, ensuring that metrics are reliable and not misleading.

Prior to training a model on a dataset, it is important to decontaminate that dataset from the desired evaluation datasets. BigCode [25] and Carper AI [26] both implement contamination detection through the use of MinHashLSH, which can be used to detect contamination prior to training a model. One concern with some modern models is that the model developers may not disclose their training data, so methods and tools have been developed to determine whether a model was trained on a specific dataset. For example canary strings, unique sequences of characters, can be included in training datasets, and Jagielski (2023) explain how to interpret canary exposure, which can identify whether a model was trained on the specified dataset. One example of canary strings can be found in the BIG-bench dataset. [27] In addition to canary strings, Shi et al. (2023) propose Min-K% probability, a method for finding possible contamination. [28]

One important note for practitioners who are performing decontamination prior to training is to see data deduplication methods for inspiration. For example, data deduplication methods that use exact matching (e.g. Bloom filters from Dolma) are also good candidates for decontamination.

4.5 Data Auditing

Auditing datasets is an essential component of dataset design. You should always spend a substantial amount of time reading through your dataset, ideally at many stages of the dataset design process. Many datasets have problems specifically because the authors did not do sufficient auditing before releasing them. The tools outlined in the data search, analysis, & exploration section ae typically sufficient to track the evolution of a dataset as it's being created. However, there are also tools that can be used to audit previously created datasets.

The Data Provenance Initiative [29] (Longpre et al., 2023b) is a good resource that documents the source, license, creator, and other metadata for over 1,800 text finetuning datasets. The Have I Been Trained? [30] tool can assist in finding and detecting data within LAION datasets. See the blog post by Jernite (2023) for more details on auditing datasets.

4.6 Review

In this section we consider, evaluate, and critically review the current state of resources for data preparation.

The community stands to benefit greatly from increased efforts on open-source data exploration tools. As we've discussed in this section, tools for data exploration and analysis are a crucial component of the iterative process of creating a dataset, however, the openly available tools for exploring data are limited. Specifically, the existing tools can search for n-grams, show random samples from the dataset, and return high-level statistics of the dataset. However, there are many additional features that may be useful. For example, after searching for an n-gram in a corpus, it may be helpful to see the containing documents to understand when this n-gram occurs and what the surrounding context is.

[24] https://github.com/allenai/dolma
[25] https://github.com/bigcode-project/bigcode-analysis/tree/main/data_analysis/decontamination
[26] https://github.com/CarperAI/decontamination/tree/main
[27] https://github.com/google/BIG-bench/.../training_on_test_set/README.md#training-on-the-test-set
[28] https://github.com/swj0419/detect-pretrain-code
[29] https://www.dataprovenance.org/
[30] https://haveibeentrained.com/

Open-source data exploration tools allow for retrospective analysis of datasets. Additionally, improving the functionality and ease-of-use for data exploration tools can help to retrospectively analyze existing datasets. For example, these tools can be used to find potential issues such as biases, illegal or copyrighted content, and personally identifiable information, as well as to ensure that they contain the advertised content (e.g. alignment datasets should contain safe text).

Consolidating on a standardized format for data storage and processing will give developers more time to focus on developing infrastructure. Furthermore, many of the data preparation tools and methods presented here exist in separate one-off repositories. This has led to limited open-sourced efforts on large-scale infrastructure. Conglomerating around a limited number of data formats can reduce friction by allowing developers to make assumptions on the format of data, thus enabling developers to focus efforts on developing scalable tooling for data preparation. For example, the use of Apache Arrow in HuggingFace Datasets (Lhoest et al., 2021b) and Numpy's `memmap` in GPT-NeoX (Andonian et al., 2021) has reduced the effort required for developers to develop memory-efficient data loading, allowing for developers to focus on building scalable tooling. Similarly, using a single data format, such as that from Dolma (Soldaini et al., 2024), can allow developers to focus on building better large-scale data infrastructure, and spend less time writing complicated code that considers multiple data formats.

Taking a fine-grained view on "high-quality" data. Referring to data as "high-quality" was originally used in the context of pretraining data, but has since become more widely adopted without a clear definition, leading to significant ambiguity in the community. In the context of cleaning and filtering web data, high-quality has referred to data that is known to have been written by humans, and has likely gone through an editing process, leading to the development of quality filters which aim to find data most similar to domains such as books or Wikipedia (Brown et al., 2020; Chowdhery et al., 2023). However, the exact definition of quality has since expanded to pretraining domains beyond web data (e.g. code, reasoning), and the phrase has been adopted in other training regimes such as preference fine-tuning. The field of foundation model development does not currently have a clear definition of what data leads to high-quality models, and the development of such definitions is a high-impact direction of research and engineering. While research on data quality for pretraining may be out of reach for smaller institutions and individual researchers, research on data quality for specific downstream use cases (e.g. code, reasoning, math) is more feasible.

For this reason, we advocate for research on data quality to be contextualized within a specific domain or set of evaluations. This will not only allow for research to progress in parallel across many groups and institutions, but we believe will also lead to definitions of quality that are much more concrete, precise, and definitive. Furthermore, the development of clear definitions lowers the barrier for creating tools that can find additional "high-quality" data.

Developing data-centric benchmarks can catalyze progress. Data-centric research is currently being done across such a wide variety of settings, and with varying goals, that it has become nearly impossible to compare methods (Albalak et al., 2024). Some recent benchmarking works have tried to address this by providing a fixed model training setup, and requiring competitors to improve the data for training. Specifically, DataComp (Gadre et al., 2023) provides a data-centric benchmark focused on image-text pairs, DataPerf (Mazumder et al., 2023) provide benchmarks for 4 settings: vision, speech, debugging, and language. Additionally for the language domain, FETA (Albalak et al., 2022) provides a benchmark for few-shot task transfer, and the Loose track from BabyLM (Warstadt et al., 2023) provides researchers a benchmark for better data selection. However, due to the large number of training settings and domains of interest, there are a wide variety of additional benchmarks and challenges that would be useful for the community (e.g. pretraining, instruction tuning, alignment, task-specific fine-tuning). Creating new data-centric benchmarks will not only allow for direct comparison between methods, improving our understanding of the data preparation methods, but also lowers the bar for entry into the field by creating an easy-to-use infrastructure, enabling increased progress.

Data preparation tools should consider not only English-centric data, but non-English and low-resource languages. While some data preparation tools may work out-of-the-box for low-resource languages (e.g. n-gram search), others may require more thought and effort (e.g. heuristic filtering). It is particularly important to include native speakers for low-resource languages throughout the data preparation process.

5 Data Documentation and Release

Documentation Best Practices

- Data documentation is essential for reproducibility, avoiding misuse, and helping downstream users build constructively on prior work.
- We recommend to start the documentation process early, as data is collected and processed.
- For datasets with multiple stakeholders or derived from community efforts, it is important to be proactive in decision-making about access, licenses, and stewardship.

5.1 Data Documentation

When releasing new data resources with a model, it is important to thoroughly document the data (Bender & Friedman, 2018; Holland et al., 2020; Gebru et al., 2021; Bommasani et al., 2024). Documentation allows users to understand its intended uses, legal restrictions, attribution, relevant contents, privacy concerns, and other limitations. An example of how to describe and document data governance decisions can be found in the BLOOM project's report (Jernite et al., 2022). The StackV2 (Lozhkov et al., 2024) is another example of a carefully curated and well documented dataset. Data documentation can also be a way to empower model trainers and downstream users of AI systems to It is common for datasets to be widely used by practitioners who may be unaware of undesirable properties (David, 2023). While many data documentation standards have been proposed, their adoption has been uneven, or when crowdsourced, as with Hugging Face Datasets, they may contain errors and omissions (Lhoest et al., 2021a; Longpre et al., 2023b).

5.2 Data Governance

Releasing all datasets involved in the development of a Foundation Model, including pretraining, fine-tuning, and evaluation data, can facilitate external scrutiny and support further research. However, releasing and hosting the data as it was used may not always be an option, especially when it includes data with external rights-holders; e.g., when data subjects' privacy, intellectual property, or other rights need to be taken into account. Proper data governance practices can be required at the curation and release stages to account for these rights.

In some jurisdictions, projects may be required to start with a Data Management Plan that requires developers to ensure that the data collection has a sufficient legal basis, follows principles of data minimization, and allows data subject to have sufficient visibility into and control over their representation in a dataset (CNIL resource sheet). Data curation steps to that end can include respecting opt-out preference signals (Spawning, HaveIBeenTrained), or applying pseudonymization or PII redaction (BigCode Governance card).

Once a dataset is released, it can be made available either broadly or with access control based on research needs (ROOTS, BigCode PII training dataset). Developers can also enable data subjects to ask for removal from the hosted version of the dataset by providing a contact address (OSCAR, PAraCrawl), possibly complemented by a membership test to check whether their data is included (Stack data portraits) or an automated process (BigCode, AmIinTheStack).

5.3 Review

Better data documentation of existing and new datasets is still needed. Most datasets still lack appropriate documentation. While data documentation tools exist they are underutilized at present. Longpre et al. (2024b) illustrate challenges in documenting provenance, consent, and authenticity of datasets at scale, and driving adoption of rigorous data standards.

Datasheets and Data Cards are just a start. While data documentation tools, such as datasheets and data cards are very useful, it is preferred to begin projects with a Data Management Plan and ensure that data collection is designed thoroughly. Considerations should include: the legal basis for collecting data, ensuring that collection is limited to the necessary data, transparency, respecting opt-out preferences and redaction of PII.

6 Model Training

Model Training Best Practices

- The foundation model life-cycle consists of several stages of training, broadly separated into pre-training and fine-tuning.
- Decisions made by developers at any stage of training can have outsized effects on the field and the model's positive and negative impacts, especially decisions made by well-resourced developers during the pre-training stage.
- Developers should be thoughtful about the effects of train-time decisions and be aware of the trade-offs and potential downstream effects prior to training.
- Due to the large economic and environmental costs incurred during model training, making appropriate use of training best practices and efficiency techniques is important in order to not waste computational or energy resources needlessly.

Thousands of tools and resources have been developed for model training. While it is not feasible to comprehensively list them, we focus our efforts on a subset of tools that are well documented, emphasize computational efficiency, or educational resources. For a more comprehensive list of training tools, we point the reader to surveys, such as the survey of foundation model training and serving systems (Zhou et al., 2024), the survey of efficient federated learning methods (Woisetschläger et al., 2024), or this "comprehensive" survey of foundation models (Zhou et al., 2023), which can be used to trace their training setup.

Foundation models are by their design frequently reused and applied for numerous diverse downstream uses. This takes the form of a multi-stage training process throughout models' lifestyles, starting with training a strong base from scratch ("pretraining") and followed by further refinement or adaptation to new use cases ("fine-tuning"). For this reason, all stages of training must be done with care: especially for pretrained models or models that will otherwise be deployed widely or used as a base for further extensions, the decisions made by model trainers at train-time can have outsized impacts on the model's characteristics, both positive and negative, or on the field as a whole.

Here, we cover a selection of existing resources for model training. We include frequently-used codebases for model training that may be useful entry points for new developers to the field, but note that we cannot cover all existing options. We additionally briefly discuss resources where practitioners can learn about improving the resource-efficiency of their training, as well as educational resources for learning about model training.

6.1 Pretraining

Model pretraining requires by far the largest amount of resources computationally during model training.

Therefore, practitioners should consider using already-optimized codebases, especially in the pretraining phase, to ensure effective use of computational resources, capital, power, and effort. Existing open-source codebases targeted at foundation model pretraining can make pretraining significantly more accessible to new practitioners and help accumulate techniques for efficiency in model training.

Various codebases may be optimal depending on the scale of model or number of devices used for pretraining. Some codebases seek to be performant while also being lightweight (OpenLM (Gururangan et al., 2023), Nanotron (nan, 2024)) for effective training at medium scales and up, while others aim for maximal performance at massive scales such as GPT-NeoX (Andonian et al., 2021), Megatron-DeepSpeed (Smith et al., 2022b), and Megatron-LM (Shoeybi et al., 2020) and have been used in practice to train language models across thousands of GPUs. Depending on the particular vision application or model architecture, good scalable codebases exist for training. For example, OpenCLIP (Ilharco et al., 2021) can be used for training CLIP text + vision models, Timm (Wightman, 2019) supports a variety of standard vision model architectures and training scripts, and UniLM supports the training of interleaved text-and-image models similar to Kosmos-2 (Peng et al., 2023). For modalities other than text or vision, tooling exists (Lhotse for audio data processing and data loading (Żelasko et al., 2021), Stable Audio Tools (AI) for training audio generation models) but is less standardized. Lastly, Karamcheti et al. (2024); McKinzie et al. (2024) explore lots of design choices for multimodal models, including training, fine-tuning, image processing strategies, and data mixing.

We emphasize the utility and importance of adopting existing codebases for pretraining, due to the difficulty of debugging and detecting errors in large-scale distributed systems as encountered in pretraining. Existing scalable codebases drastically reduce the chances of silent failures or correctness issues lowering the end model quality.

6.2 Fine-tuning

Fine-tuning, or other types of adaptation performed on foundation models after pretraining, are an equally important and complex step in model development and have the ability to significantly steer the behaviors and characteristics of the end model. Fine-tuned models are also more frequently deployed than base models, making their safety and usability very important. Here we discuss a subset of finetuning resources that are well documented, flexible, and some of which cater to computational efficiency.

While fine-tuning is significantly less resource-intensive than pretraining, there are still many relevant considerations for practitioners. Use of widely adopted tools such as libraries designed for fine-tuning (Axolotl (OpenAccess-AI-Collective), trlX (Havrilla et al., 2023), Levanter (`https://github.com/stanford-crfm/levanter`)) can ensure greater ecosystem compatibility of resulting models, or reduce the barrier to experimentation by abstracting away common pitfalls or providing guidance on effective hyperparameters. Similarly, the use of techniques such as QLoRA (Dettmers et al., 2023) or other popular parameter-efficient fine-tuning approaches and libraries (peft (Mangrulkar et al., 2022) , Otter, LLaMA-Adapter, LLaVA (Li et al., 2023b; Zhang et al., 2023a; Liu et al., 2023a)) can allow for fine-tuning for lower cost or on more accessible hardware.

6.3 Efficiency and Resource Allocation

Knowledge of training best practices and efficiency techniques can reduce costs to train a desired model significantly. Beyond systems-level optimizations and approaches to increase efficiency, the most important technique is determining the most efficient *allocation* of resources, such as allocating compute between model size and dataset size for a given budget for the best results. The most common approach to solving this problem is to apply *scaling laws* (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023b), a common tool for cheaply extrapolating findings across scales of cost in order to make decisions based on smaller-scale experiments for a final model training run.

Additionally, practitioners seeking to embrace an open approach to model development should consider how their decisions when training a foundation model may have impacts long after that model's creation and release. For instance, a model that is released openly but is too computationally demanding to be run on consumer-grade hardware will be limited in its impact on the field, or a model trained to minimize training compute but not minimize inference cost may result in a greater environmental impact than spending more training compute in the first place for a cheaper-to-infer model (Hoffmann et al., 2022; Touvron et al., 2023). Practitioners should thus be aware of potential second-order effects of their model releases and training choices.

6.4 Educational Resources

Training models at any scale can be quite daunting to newer practitioners. However, there are options available for learning about how to train and run foundation models.

Resources which themselves collate and provide reading lists or recommendations for learning about large-scale ML training (EleutherAI Cookbook (Anthony et al., 2024), ML Engineering Open Book (Bekman)) can be an especially useful place to begin. Additionally, minimal codebases created as educational examples such as NanoGPT (Karpathy, 2023), or other blog posts detailing fundamental concepts in training or running foundation models may be useful (Chen, 2022; Anthony et al., 2023). We recommend that practitioners review these resources and use them to guide further reading about model training and usage.

6.5 Recommendations

Regardless of the stage or cost of training being performed, we urge developers to carefully consider the impacts of their training choices, and how they can be improved, as we have described. We hope the linked resources will be a helpful jumping-off point.

We additionally encourage practitioners to use the codebases linked as a foundation for their work, to avoid "reinventing the wheel" for every new project. Unifying and collectively improving existing codebases, tooling, and commonly used standards around model training, as with tooling around other parts of the foundation model development process, can allow for the entire community to benefit from others' effort via using codebases that have been well-tested for correctness and scalability. This can leverage the unique advantages of open model development via strength in numbers.

6.6 Review

More resources, especially non-English ones, for lowering the barrier to entry for less technical developers are needed. Current tooling requires at minimum a technically knowledgeable user or proficient programmer, and further may have lacking documentation or be high in complexity. Openly available tools for training, especially fine-tuning, models that do not presume familiarity with training methods or even programming would have the potential to allow many more users not formally educated in computer science to customize and build models suited to their use cases, such as for lower-resource languages.

More standardized and centralized resources, especially cross-modality, should be focused on in the future. Many current resources are centered around language model training. However, resources for other modalities are often more scattered or rely on custom architectures and implementations. Exploring topics such as multimodal fine-tuning best practices, and multimodal preference learning, is an important future area. For instance, there is a lack of efficient tooling for data loading for multimodal training. While WebDataset[31] offers a resource for images, there are fewer options for videos, which can cause the GPU to stay idling, waiting for data. Future work should look to unify these techniques and modalities in the same tools and infrastructure to enable more and easier investigation or transfer of techniques that work on language models to the rest of the field. Even within the text modality, an *interoperable* ecosystem with shared, well-tested code building blocks and standards should be pursued, to provide a strong base for further extension and the pooling of developer efforts.

Other options for massively collaborative development should be pursued. To allow for better pooling of resources, both over time and across the community, methods for more collaborative training are an important future frontier. Some of these methods are in their early stages already being provided by tools such as MergeKit (Goddard et al., 2024) and being explored by the research community (Matena & Raffel, 2022; Don-Yehiya et al., 2023; Stoica et al., 2024; Yadav et al., 2023). By enabling better re-use of existing artifacts or allowing a greater number of collaborators to contribute to building a model together, these methods could potentially make model training as parallelizable and reusable as work on data improvement, and the merging of efforts and compute expended could allow a wider range of contributors to steer model development.

7 Environmental Impact

Environmental Impact Best Practices

- Training and deploying AI models impacts the environment in several ways, from the rare earth minerals used for manufacturing GPUs to the water used for cooling datacenters and the greenhouse gasses (GHG) emitted by generating the energy needed to power training and inference.
- Developers should report energy consumption and carbon emissions separately to enable an apples-to-apples comparisons of models trained using different energy sources.
- It is important to estimate and report the environmental impact not just of the final training run, but also the many experiments, evaluation, and expected downstream uses.
- It is recommended, especially for major model releases, to measure and report their environmental impact, such as carbon footprint, via mechanisms such as model cards (see Section 5).

[31] https://github.com/webdataset/webdataset

7.1 Estimating Environmental Impact

The environmental impact of AI is of increasing concern (Schwartz et al., 2020). Estimating the environmental impact of model development is a challenging task due to the number of relevant, hard to compute, and often unrecorded variables. For instance, to estimate the Life Cycle Assessment (LCA) of model development (Klöpffer, 1997), variables include: model size, architecture, duration of training, number of training runs, storing and transferring data, hardware manufacturing, the specific type and setup of the hardware, network configuration, the geographic location of the data center, the carbon intensity of the energy grid, the power usage effectiveness (PUE) of cooling, overhead, and the broader data center infrastructure, as well as a similar set of questions for various deployment/inference settings (Patterson et al., 2021). Many of these variables are also infeasible to precisely estimate, given for instance, that Nvidia does not disclose the carbon footprint of its GPUs (Luccioni et al., 2023b). Another challenge includes the variety of relevant environmental outcome measures, from CO2 emissions, to energy footprint, or water use.

Existing tools to measure environmental impact rely on a series of assumptions and estimates based on the available information. Perhaps the most accurate tools are those built into cloud services, which are able to trace the hardware configurations and geographical locations of data centers but may not otherwise be publicly available to users. The Azure Emissions Impact Dashboard (`https://www.microsoft.com/en-us/sustainability/emissions-impact-dashboard`), AWS carbon footprint tool (`https://aws.amazon.com/aws-cost-management/aws-customer-carbon-footprint-tool/`), and Google Cloud Carbon Footprint measurement system (`https://cloud.google.com/carbon-footprint?hl=en`) are three such systems, available only when using their cloud services directly. Independent systems, such as Carbontracker (Anthony et al., 2020), CodeCarbon (Schmidt et al., 2021), or the Experimental Impact Tracker (Henderson et al., 2020) offer basic estimate modeling and reporting, based on limited input information. Similarly, Li et al. (2023d) provides an estimate tool for the water usage footprint of language model training and deployment. ML CO2 Impact (`https://mlco2.github.io/impact/`) improves these repositories with a wrapper interface for easier use. However, these services are forced to trade-off between ease of use and accuracy, as significant input information is required to obtain precise results, which imposes a burden on users and widespread adoption.

7.2 Effective use of resources

Several decisions made prior to model training can have significant impacts on the upstream and downstream environmental impact of a given model. Empirical scaling laws can be used to find the best allocation of resources. Kaplan et al. (2020); Hoffmann et al. (2022) estimate the optimal model size and training duration, given a training compute budget. And Aghajanyan et al. (2023) investigates the equivalent efficient compute allocation for multi-modal settings. When working with text training data that is constrained, recent work explores how to allocate compute efficiently (Muennighoff et al., 2023b). For models frequently used downstream, it is important to consider the inference footprint and inference cost during model creation (Gadre et al., 2024b), to minimize the environmental impact of inference. For further resources and discussion, see 6.3.

7.3 Review

In this section we critically review the current state of resources for environmental impact analysis. First, we note a dearth of transparency, from hardware manufacturers, data centers, and corporate model developers stymies environmental impact estimates, and particularly product comparisons for users. Lastly, scaling laws research needs to expand to cover newer multi-modal modeling efforts, as well as provide intuitive tooling for open developers to adopt.

Transparency from consumer-level API providers, into query-level energy and environmental usage measures. Currently, some of the most widely adopted generative AI services, including the APIs and playgrounds from OpenAI, Google, Anthropic, Inflection, Midjourney, and others, do not expose any information into the environmental footprint of using their models (Bommasani et al., 2023a). As these systems dominate consumer market usage (Korinek & Vipra, 2023), this leaves a wide gap in our knowledge of net effects. The scientific community relies largely on assumptions and ballpark estimates, both for training and inference impact.

Transparency from hardware makers and data centers, into fine-grained energy and environmental usage measures. Upstream of AI developers, the data centers and hardware makers expose limited information about environmental

and energy measures (Luccioni et al., 2023b). These metrics would enable more accurate and real-time estimates of compute, for closed and open developers.

"Energy Star" standards for non-technical users to fairly compare AI services. Data center, hardware, and developer transparency are a precursor to accurate estimates and, more importantly, *environmental footprint competition* between corporate services. Currently, these services compete on model quality and price, but not on environmental impact—a property that many consumers are likely to care about if given the option. "Energy Star" standards, from other industries, allow competitors to claim equal or better services, at less energy expenditure (Brown et al., 2002). These sorts of apples-to-apples comparisons may be necessary to inhibit negative environmental consequences from AI.

Scaling laws research currently lacks empirical evidence for the new wave of multi-modal models, and intuitive user interfaces for new developers. Scaling laws research has focused predominantly on text (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023b; Gadre et al., 2024b), with limited work for multi-modal foundation models (Aghajanyan et al., 2023). As developers increasingly pursue image, video, and speech models (both in input and output), such as Sora (Brooks et al., 2024), Stable Video Diffusion (Blattmann et al., 2023), Claude 3 (Anthropic, 2024b), and Whisper (Radford et al., 2023), the efficient scaling laws are currently under-investigated. Secondly, while scaling law research investigates fundamental questions, it can be presented in unapproachable, complex ways. The ecosystem lacks tools for less technical developers to heed efficient compute estimates, such as a plug-and-play interface.

8 Model Evaluation

Model Evaluation Best Practices

- Model evaluation is an essential component of machine learning research. However many machine learning papers use evaluations that are **not reproducible or comparable to other work**.
- One of the biggest causes of irreproducibility is failure to report prompts and other essential components of evaluation protocols. This would not be a problem if researchers released evaluation code and exact prompts, but many prominent labs (OpenAI, Anthropic, Meta) have not done so for model releases. When using evaluation results from a paper that does not release its evaluation code, **reproduce the evaluations using a public codebase**.
- Examples of high-quality documentation practices for model evaluations can be found in Brown et al. (2020) (for bespoke evaluations) and Black et al. (2022); Scao et al. (2022); Biderman et al. (2023) (for evaluation using a public codebase).
- Expect a released model to be used in unexpected ways. Accordingly, try to evaluate the model on benchmarks that are most related to its prescribed use case, but also its failure modes or potential misuses.
- All evaluations come with limitations. Be careful to assess and communicate these limitations when reporting results, to avoid overconfidence in model capabilities.

8.1 Capabilities

Many modern foundation models are released with general conversational abilities, such that their use cases are poorly specified and open-ended. This poses significant challenges to evaluation benchmarks that are unable to critically evaluate so many tasks, applications, and risks fairly or systematically. As a result, it is important to carefully scope the original intentions for the model, and to tie evaluations to those intentions. Even then, the most relevant evaluation benchmarks may not align with real use, and so developers should qualify their results, and carefully supplement them with data from real user/human evaluation settings where feasible.

For language models, common capabilities benchmarks include those that evaluate models on narrow tasks such as software engineering (Jimenez et al., 2023), topic classification (Adelani et al., 2023), and explaining code (Muennighoff et al., 2023a). More comprehensive evaluation suites, such as the Language Model Evaluation Harness (Gao et al., 2023a) and HELM (Liang et al., 2023), are also common. Leaderboards like LMSys Chatbot Arena (Zheng et al., 2023) offer another type of capability evaluation based on human feedback.

There are far fewer capability evaluations for other modalities. Comprehensive evaluation suites exist for vision models as well (Lee et al., 2023b; Awadalla et al., 2023), but they are relatively less well developed. There are also a number of common benchmarks for evaluating vision models on a large number of tasks (Gadre et al., 2023; Liu et al., 2023b; Fu et al., 2023). Evaluations for speech models' capabilities are still nascent, with the OpenASR Leaderboard,[32] which ranks models based on their Word Error Rate and Real-Time Factor, and the Edinburgh International Accents of English Corpus (Sanabria et al., 2023) as leading examples.

The cheatsheet includes common benchmarks as of December 2023, but we caution that each comes with substantial limitations. For instance, many benchmarks based on multiple choice exams are not indicative of real user questions, and can be gamed with pseudo-data contamination. Additionally, while leaderboards are exceedingly popular, model responses are often scored by other models, which have implicit biases to model responses that are longer, and look similar to their own (Dubois et al., 2023).

8.2 Harm & Hazard Taxonomies

Taxonomies provide a way of categorising, defining and understanding risks and hazards created through the use and deployment of AI systems. Some taxonomies focus primarily on the types of interactions with models that *create* a risk of harm (often called "hazards") whereas others focus on the negative effects that they lead to (often called "harms"). Some taxonomies focus on existing issues, such as models that create hate speech or child abuse material, as well as more intangible or indirect forms of harm, such as the risk that models perpetuate biases and stereotypes, and misrepresent social groups. Other taxonomies are focused on the longer-term threats posed by more sophisticated models, such as ultra-personalised disinformation, cybersecurity threats, and military use (Brundage et al., 2018). Some work has also focused on categorising catastrophic or "existential" risks presented by Artificial General Intelligence, such as rogue AI agents and Chemical, Biological, Radiological, Nuclear and high-yield Explosive weapons (Carlsmith, 2022; Hendrycks et al., 2023; Bucknall & Dori-Hacohen, 2022). Further, a few taxonomies also assess the available evidence for the risks and hazards, discuss their impact, and offer mitigation strategies (Deng et al., 2023; Kapoor et al., 2024; Klyman, 2024).

Many taxonomies are released with an associated dataset, which can be used to either train models to minimise safety risks or to evaluate those risks. Several datasets have been released as benchmarks (e.g. Wang et al., 2024), which can be used to track progress across the community. We provide a non-exhaustive list of existing taxonomies with datasets that are available open-source. We also note forthcoming work planned by organisations like ML Commons, which aims to standardise assessment of AI safety risks by introducing a new benchmark, comprising a taxonomy and dataset.[33].

1. **TrustLLM** is a benchmark that covers six dimensions in English, including truthfulness, safety, fairness, robustness, privacy, and machine ethics (Sun et al., 2024). The benchmark comprises over 30 datasets from existing research. In the paper, they test 16 open-source and proprietary models, and identify critical safety weaknesses.

2. **SafetyBench** is a benchmark that covers eight categories of safety, in both English and Chinese (Zhang et al., 2023b). Categories include Offensiveness; Unfairness and Bias; Physical Health; Mental Health; Illegal Activities; Ethics and Morality; and Privacy and Property. Unlike most safety evaluation datasets, SafetyBench comprises multiple choice questions which makes automated evaluation of models far easier. In the paper, they test 25 models and find that GPT-4 consistently performs best.

3. **DecodingTrust** is a benchmark that covers eight dimensions of AI safety and trustworthiness in English (Wang et al., 2024). It covers a range of safety criteria such as Toxicity; Stereotypes; Adversarial Robustness; Out-of-Distribution Robustness; Privacy; Machine Ethics; And Fairness. The benchmark has a leaderboard that is hosted on HuggingFace.

4. **HarmBench** is a standardized evaluation framework for automated redteaming of LLMs in English (Mazeika et al., 2024). It covers 18 widely used red teaming methods, such as Persona, stochastic-few shot, PEZ and GBDA. The benchmark has been designed with both seven semantic categories (e.g. Cybercrime,

[32] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard
[33] https://mlcommons.org/working-groups/ai-safety/ai-safety/

Misinformation and Bioweapons) and four "functional categories" (e.g. Standard behaviours). In the paper, 33 LLMs are tested against HarmBench.

5. **BigBench** (Srivastava et al., 2023b) and HELM (Liang et al., 2023) contain tests that are related to safety, such as toxicity, bias and truthfulness in BigBench and toxicity, bias, disinformation, copyright infringement and truthfulness in HELM. Both make use of the widely-used RealToxicityPrompts dataset (Gehman et al., 2020a). Terms such as "toxicity" and "offensiveness" have been criticised in some papers for being overly broad and easy to misinterpret (Vidgen et al., 2019), and more recent work has tended to use more fine grained terms.

6. Individual datasets have also been released that can be used to assess specific safety risks of models, such as SimpleSafetyTests which tests for clear-cut safety problems (Vidgen et al., 2023) and XSTest which tests for model false refusal (Röttger et al., 2024).

7. **SafetyPrompts** is a website that hosts datasets for evaluating the safety of models [34] It does not aggregate or combine the datasets that it hosts, but has a basic review process to check the quality and integrity of the datasets.

Almost all of the taxonomies described here are informed by practitioners' experiences of tackling safety issues in AI models; prior research; and exploratory red teaming. Many have drawn heavily on existing work in social media trust and safety, such as Banko et al. (2020); Dev et al. (2022). They are all top-down in the sense that they define categories of hazard (or harm) and then find, create or curate prompts that match those categories. An alternative way of addressing AI safety is to start from the bottom up and to task red teamers with creating their own categories. Grounded-theory style approaches in linguistics can be used to then standardise the categories and ascribe some structure to the taxonomy. Across all methods of creating taxonomies (and datasets), there is a substantial focus on text-only models and future work should pay more attention to multimodal and non-text based models. Similarly, there is a strong bias towards English language and the Western cultural context. This should also be addressed in future work.

8.3 Risks & Harms

Model developers have used various techniques to address risks and mitigate harms. Broadly, these attempts fall into three categories, roughly ordered by effectiveness.

In-context learning. In-context learning can be used to add instructions to a model's system prompt to steer the model's outputs. For example, OpenAI's DALL-E 3 model included instructions to avoid outputting copyrighted characters[35]. In some sense, this is the easiest model-level intervention: developers do not need to retrain or fine tune the model, neither do they need to rely on external API calls, such as to third-party content moderation endpoints. However, this comes at a cost: system prompts are easy to jailbreak, and as a result, interventions based on in-context learning might be more brittle compared to other guardrails.

Model alignment. One of the most prominent approaches for mitigating risks and harms is aligning models with preference data. This usually involves supervised fine tuning or training reward models that steer the outputs of language models. Popular techniques include reinforcement learning with human (RLHF) and AI (RLAIF) feedback (Bai et al., 2022; Ouyang et al., 2022). To help end users understand the performance of reward models, Lambert et al. (2024) develop a leaderboard for evaluating reward models for various desiderata, including safety. Still, model alignment techniques can be brittle against simple modifications to the model, such as fine tuning (Qi et al., 2023), even when the model is fine tuned using benign data (He et al., 2024).

Guardrails on model inputs and outputs. The two interventions above rely on changing the model behavior via in-context learning or fine tuning. However, guardrails on model inputs and outputs that lie *outside* the model might be a more robust intervention for preventing harmful content generation. For example, several model developers provide moderation endpoints that can be used to filter inappropriate user requests or model outputs. These can be general-purpose endpoints that can be modified for filtering and moderation (e.g., Cohere's classification endpoint

[34] `https://safetyprompts.com/`

[35] See: `https://the-decoder.com/dall-e-3s-system-prompt-reveals-openais-rules-for-generative-image-ai/`

modified for toxicity detection[36] or Anthropic's suggested prompt for content moderation[37]) as well as endpoints that are specifically built for content moderation (e.g., Perspective API (Lees et al., 2022) and OpenAI's moderation endpoint [38]). Google's PaLM and Gemini models allow API users to set thresholds based on the safety likelihood and severity of model outputs[39]).

The examples above are all from closed model developers. Recently, several open models have also been proposed for moderating model inputs and outputs. For example, Llama-Guard by Meta (Inan et al., 2023) can be used to filter harmful content. Nvidia's NEMO guardrails allow model providers to add programmatic guardrails to filter content (Rebedea et al., 2023).

8.4 Review

A shift away from evaluating models, and towards evaluating systems. Existing evaluation practices and reporting often focus on probing individual models, in unconstrained settings. However, deployed systems often operate in the context of multiple interactive models, moderation endpoints, sophisticated decoding strategies, and rule-based constraints, that make up a "system". More pointedly, for dual-use systems, the context of use *outside the system* is what defines the presence of harm: Narayanan & Kapoor (2024) argue that "defenses against misuse must primarily be located outside models". Prior work has emphasized the importance of the system (Dobbe, 2022) and context (Raji & Dobbe, 2023) in diagnosing and resolving AI safety challenges. As a result, efforts to evaluate models alone are inherently limited, both in their findings, and informing effective changes. Safety problems in particular are better addressed by evaluations that consider the context (e.g. attack vector, deployed use setting), as well as the interactions between elements of a foundation model system (of which only one is a the model itself). However, these types of evaluations can be difficult in the absence of transparency into the components of deployed systems (Bommasani et al., 2023a). Recent work has even shown that independent researchers can face significant obstacles in fairly evaluating proprietary systems (Longpre et al., 2024a).

A shift away from evaluating on static toxicity/safety benchmarks, and towards evaluating on real-world, dynamically changing naturalistic observations, as well as human-interaction studies. Existing widely used model risk and safety evaluations, such as RealToxicityPrompts (Gehman et al., 2020b), or bias benchmarks such as BBQ (Parrish et al., 2022), BOLD (Dhamala et al., 2021), and HolisticBias (Smith et al., 2022a), have been effective at testing models' superficial bias tendencies. However, these evaluations are static, not grounded in real-world contexts, and pay no heed to the human-interaction element—*i.e. the actual system users*. These limitations suggest several dimensions of improvement for future safety benchmarks:

- **Naturalistic observations** Collecting natural interactions with users enable researchers to identify realistic tasks and prompts, to simulate real harms. For instance, WildChat (Zhao et al., 2023b) collects voluntary user interactions with OpenAI systems through proxy interfaces—however these datasets may be heavily skewed by the types of users that adopt it.

- **Domain expert designed tasks** LegalBench (Guha et al., 2024) offers an alternative naturalistic design, whereby evaluations are hand-crafted by the practitioners (in this case, legal professionals) in the field of evaluation. These benchmarks distinguish themselves from existing evaluation suites in their relevance to natural, real-world usage.

- **Human-interaction studies** Most evaluations are non-interactive—they target the model, without real-world users. This form of analysis can identify model flaws, but falls short of investigating the actual affects, harms, and interaction patterns on users. Lee et al. (2023a) proposes a framework to evaluate interactive user experience, without which notions of harm and safety are under-developed. Le Ferrand et al. (2022) illustrates the importance of interactive evaluation, particularly on non-Western users.[40]

[36]See: `https://docs.cohere.com/reference/toxicity-detection`

[37]See: `https://docs.anthropic.com/claude/docs/content-moderation`

[38]See: `https://platform.openai.com/docs/guides/moderation/quickstart`

[39]See: `https://cloud.google.com/vertex-ai/generative-ai/docs/configure-safety-attributes-palm`, `https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/configure-safety-attributes`

[40]In this study, Australian Aboriginal use of an information retrieval app illuminated false assumptions and expectations in the evaluation procedure.

These three evaluation dimensions enable more realistic studies of user interaction with foundation model systems. Collecting naturalistic or interactive data can also provide *continuous and dynamic* data sources, which are particularly beneficial when there is rapid model development (or the habit of training on prior evaluation sets). More generally, research into dynamic and evolving benchmarks, that test beyond the training set distribution, are an important research direction (Yu et al., 2023; Kiela et al., 2021).

A shift away from reporting evaluations, to releasing reproducible evaluation scripts. There are dozens of choices that affect the results of an evaluation (Anthropic, 2023). Some choices include, but are not limited to:

- **Prompt format** Results vary dramatically depending on if the input prompts are zero-shot, few-shot, chain-of-thought, and also if the prompts were manually refined directly on the test set. Multiple choice benchmarks have both versions with and without the answer choices given in the prompt. (See MMLU as a widely used dataset with inconsistencies in use and standardization.[41]

- **Decoding strategies** The decoding algorithm and its hyperparameter choices, such as the temperature and sampling probabilities, will affect model behavior. For systems, rather than models, responses can be the result of multiple iterations or models, or rely on external tools or sources. Ensembling techniques like self consistency (Wang et al., 2022) also improve performance significantly.

- **Evaluation metric** The choice of evaluation metric can impact the apparent magnitude of differences between models, and even their ranking (Schaeffer et al., 2024).

- **Human review setup** For human preference evaluations, several details affect fair evaluation: whether the response selection is sufficiently model-blind, the attentiveness and expertise of the annotators to the given topics, and the chosen rubric. Human preferences can also be skewed by the same factors as model-based evaluations (Hosking et al., 2024; Xu et al., 2023; Wu & Aji, 2023)

Without transparency and reproducibility integrated into the evaluation procedure, the axes of design freedom can allow developers to game results, and prevent fair, apples-to-apples comparisons. For these reasons, only evaluation scripts that are directly executable by third-parties provide **verifiable reproducibility**. Executable evaluation scripts also allow auditors to unpack the choices made in evaluation. Auditors can quickly experiment with different prompt formats, decoding parameters, and evaluation metrics, to shed light on the scientific veracity of the claims. While evaluation documentation is helpful, the required breadth of information inevitably leads to subtle omissions, and even if the information is comprehensive, it does not provide verifiable reproducibility. For these reasons, we believe the evidence is clear that AI safety reporting is not sufficiently reliable or trustworthy without verifiable execution scripts.

Ground harm and hazard taxonomies in empirical observations. Existing taxonomies of harm are often created to cluster existing safety benchmarks, which are mostly detached from real observations (Sun et al., 2024; Zhang et al., 2023b; Hendrycks et al., 2023). As the taxonomies of harm can guide practitioners priorities, this could lead to neglected or over-emphasized areas of safety research. It is essential that future harm taxonomies are strongly grounded in empirical or naturalistic observations, through research conducted with *real users*, rather than hypothetical situations envisioned by researchers.

Extending risks and harms studies to multimodal and highly sensitive attacks. A great deal of research into risks and harms is focused exclusively on text, in English, and on *more conservative* safety risks such as toxicity and bias (Gehman et al., 2020a; Hartvigsen et al., 2022). However, recent work has emphasized the particular risks of generative image, speech, and video models, being used to create deepfakes, NCII or CSAM (Kapoor et al., 2024; Thiel et al., 2023b; Lakatos, 2023a). Other work has illustrated the much greater efficacy of jailbreaks and attacks in non-English languages, as compared to English (Yong et al., 2023). Multimodal jailbreaking is an emerging research area, with some nascent work (Qi et al., 2024; Shayegani et al., 2023; Niu et al., 2024). Research into risks and harms from autonomous weapons systems (AWS) is also highly under-explored, and an emerging risk (Simmons-Edler et al., 2024; Longpre et al., 2022).

[41] https://crfm.stanford.edu/2024/05/01/helm-mmlu.html

9 Model Release & Monitoring

Model Release & Monitoring Best Practices

- Release models with accompanying, easy-to-run code for inference, and ideally training and evaluation.
- Document models thoroughly to the extent possible. Model documentation is critical to avoiding misuse and harms, as well as enabling developers to effectively build on your work.
- Open source is a technical term and standard with a widely accepted definition that is maintained by the Open Source Initiative (OSI) (Initiative, 2024). Not all models that are downloadable or that have publicly available weights and datasets are open-source; open-source models are those that are released under a license that adheres to the OSI standard.
- The extent to which "responsible use licenses" are legally enforceable is unclear. While licenses that restrict end use of models may prevent commercial entities from engaging in out-of-scope uses, they are better viewed as tools for establishing norms rather than binding contracts.
- Choosing the right license for an open-access model can be difficult. Apache 2.0 is the most common open-source license, while responsible AI licenses with use restrictions have seen growing adoption. Consider using one of the available tools for selecting the right open-source license for your model artifacts.
- Frameworks for monitoring and shaping model usage have become more prevalent as policymakers have attempted to constrain certain end uses of foundation models. Several approaches include adverse event reporting, watermarking, and restricting access to models in limited ways. Consider providing guidance to users on how to use your models responsibly and openly stating the norms you hope will shape model use.

9.1 Model Documentation

When models, code or applications are released, whether openly or not, it is important that they are documented thoroughly. Documentation should specify how to use the model, recommended and non-recommended use cases, potential harms, state or justify decisions made during training, and more. Documenting models is important not just for responsible development, but also to enable other developers to effectively build on a model. Models are not nearly as useful as artifacts if not properly documented. Model cards (Mitchell et al., 2019) are widely adopted standard for documenting models. Several tools have been developed that support the creation of model cards[42].

9.2 Reproducibility

Code to reproduce results are an important complement to other forms of documentation Kapoor et al. (2023). Releasing assets that reproduce results mean that scientific claims can be verified, and that systems can be interrogated, tested and audited Missing, incomplete, or poorly documented code hinders progress. There are tools that help make model training, inference and evaluation reproducible. Anaconda and Docker make necessary environments and dependencies easier to manage. Google Colab and Jupyter notebooks enable easily shareable code snippets and organized tutorials. The Language Model Evaluation Harness provides a framework for prompting and testing generative language models on a large number of different evaluation tasks (Gao et al., 2023a).

9.3 Licensing

Licensing is a mechanism creating a legally enforceable agreement that governs use of artifacts. Licenses with use restrictions can be used to limit the ability of certain categories of stakeholders to re-use or adapt the models (Contractor et al., 2022; Foundation, 2024). The MIT and Apache 2.0 licenses are the most commonly used. Responsible AI Licenses, including BigScience's Open RAIL, have seen growing adoption. However these also face criticism around how they pose challenges for well-intentioned actors and because their enforceability remains an open question (Downing, 2023). While RAIL licenses that restrict end use of models may prevent commercial entities from engaging in out-of-scope uses, they are better viewed as tools for establishing norms rather than binding contracts.

[42] https://huggingface.co/blog/model-cards

9.4 Usage Monitoring

Some open foundation model developers attempt to monitor the usage of their models, whether by watermarking model outputs or gating access to the model. The cheatsheet provides resources related to usage monitoring, including examples of how to watermark content, guidance on appropriate use, guidance on reporting adverse events associated with model use[43], and ways to limit some forms of access to models. Several of these approaches have significant drawbacks: for example, there are no known robust watermarking techniques for language models and there are limits to watermarking for image models (Kirchenbauer et al., 2023; Saberi et al., 2023). As with many of the sections above, usage monitoring remains an area of active research.

9.5 Recommendations

Models should be released with accompanying documentation *and* easy-to-run code for training, evaluation and inference. Document model thoroughly to the extent possible. These are critical to avoiding misuse and harms and enabling developers to effectively build on your work. A well documented environment, code, and versions of the appropriate datasets.

Be thoughtful about the type of license to use for artifacts. Open source is a technical term and standard with a widely accepted definition (Initiative, 2024). If there is a risk of misuse then consider behavioral restrictions from a standardized tool (McDuff et al., 2024).

Frameworks for monitoring and shaping model usage have become more prevalent as policymakers have attempted to constrain certain end uses of foundation models. Several approaches include adverse event reporting, watermarking, and restricting access to models in limited ways. Consider providing guidance to users on how to use your models responsibly and openly stating the norms you hope will shape model use.

9.6 Review

In this section we critically review the current state of resources for model release and monitoring, from our survey.

Reproducibility, especially through executable code, benefits all parties. The research community has benefited substantially from openness and transparency in how foundation model artifacts are documented and released. As have proprietary developers who benefit from the rapid prototyping and innovations on their released technical systems, propelled by the open community. Our review of resources suggests that often the most widely adopted tools are those that are not just well documented/described, but also those that are easy to run with executable code.

Usage monitoring remains challenging, and offers both advantages and disadvantages. Existing tools such as watermarking for AI outputs (Kirchenbauer et al., 2023; Saberi et al., 2023) or model fingerprinting (Xu et al., 2024) offer some degree of verification regarding the source of data or models. However, their robustness to adversarial removal, or to detection, remain dubious—this is particularly true for text data. This can result in over-confidence that these methods provide a panacea, and instead result in false positives. There are also open questions on privacy as to the right for individuals to produce content without attribution. We hesitate to prescribe these nascent solutions broadly, without fully understanding the particular context under consideration for the data, the model, and their potential uses and abuses.

Permissions and restrictions around model use should be more explicit about their data provenance, and the other upstream ingredients which may impact use intentions, consent, and permissions. The ethical, responsible or legal use of a model may depend on a series of upstream factors, beyond the final developers' license. This could include the license of the datasets used for training, the terms of service attached to those data sources (e.g. if some data is synthetically generated), or the license of the base model (if finetuned). The norms, best practices, and legal relevance of each of these factors is evolving, jurisdiction-dependent, and beyond the scope of this cheatsheet. However, we suggest developers document these factors in their model releases, beyond their own license, such that downstream developers and users can make informed decisions.

[43] E.g. `https://www.microsoft.com/en-us/photodna`

References

Nanotron. https://github.com/huggingface/nanotron, 2024.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*, 2023.

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.

Stability AI. Stable audio tools. Github Repo. URL https://github.com/Stability-AI/stable-audio-tools.

Alon Albalak, Yi-Lin Tuan, Pegah Jandaghi, Connor Pryor, Luke Yoffe, Deepak Ramachandran, Lise Getoor, Jay Pujara, and William Yang Wang. FETA: A benchmark for few-sample task transfer in open-domain dialogue. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10936–10953, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.751. URL https://aclanthology.org/2022.emnlp-main.751.

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*, 2023.

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.

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudeker, Giulia Osti, Daniel van Strien, et al. Datasheets for digital cultural heritage datasets. *JOURNAL OF OPEN HUMANITIES DATA*, 9(17):1–11, 2023.

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.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. GPT-NeoX: Large scale autoregressive language modeling in PyTorch, 8 2021. URL https://www.github.com/eleutherai/gpt-neox.

Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

Quentin Anthony, Stella Biderman, and Hailey Schoelkopf. Transformer math 101. GitHub Repo, 2023. URL blog.eleuther.ai/.

Quentin Anthony, Hailey Schoelkopf, and Stella Biderman. The EleutherAI Model Training Cookbook. GitHub Repo, 2024. URL https://github.com/EleutherAI/cookbook.

Anthropic. Challenges in evaluating ai systems, October 2023. URL https://www.anthropic.com/news/evaluating-ai-systems.

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 3 2024a. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024b. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

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

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*, 2023.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pp. 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

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

Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 125–137, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.16. URL https://aclanthology.org/2020.alw-1.16.

Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022.

Stas Bekman. Machine learning engineering open book. GitHub Repo. URL https://github.com/stas00/ml-engineering.

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.

Manish Bhatt, Sahana Chennabasappa, Cyrus Nikolaidis, Shengye Wan, Ivan Evtimov, Dominik Gabi, Daniel Song, Faizan Ahmad, Cornelius Aschermann, Lorenzo Fontana, et al. Purple llama cyberseceval: A secure coding benchmark for language models. *arXiv preprint arXiv:2312.04724*, 2023.

Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*, 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024b.

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, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023a.

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*, 2023b.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

Rishi Bommasani, Dilara Soylu, Thomas I Liao, Kathleen A Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models. *arXiv preprint arXiv:2303.15772*, 2023b.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports. *arXiv preprint arXiv:2402.16268*, 2024.

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`.

R Brown, C Webber, and JG Koomey. Status and future directions of the energy star program. *Energy*, 5(27):505–520, 2002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018.

Benjamin S. Bucknall and Shiri Dori-Hacohen. Current and near-term ai as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 119–129, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534146. URL `https://doi.org/10.1145/3514094.3534146`.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.75. URL `http://dx.doi.org/10.18653/v1/2023.acl-short.75`.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, et al. Nusacrowd: Open source initiative for indonesian nlp resources.

Joseph Carlsmith. Is power-seeking ai an existential risk?, 2022.

Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. Reclaiming the digital commons: A public data trust for training data. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 855–868. Association for Computing Machinery, 2023. doi: 10.1145/3600211.3604658. URL `https://doi.org/10.1145/3600211.3604658`.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Carol Chen. Transformer inference arithmetic. `https://kipp.ly/blog/transformer-inference-arithmetic/`, 2022.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 778–788, 2022.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jiawen Deng, Jiale Cheng, Hao Sun, Zhexin Zhang, and Minlie Huang. Towards safer generative language models: A survey on safety risks, evaluations, and improvements, 2023.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. Assessing language model deployment with risk cards, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. On measures of biases and harms in NLP. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pp. 246–267, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

Mark Dingemanse and Andreas Liesenfeld. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp. 5614–5633, 2022.

Roel Dobbe. System safety and artificial intelligence. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1584–1584, 2022.

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.

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. Cold fusion: Collaborative descent for distributed multitask finetuning, 2023.

Kate Downing. Ai licensing can't balance "open" with "responsible". The Law Office of Kate Downing's Blog, 2023. URL https://katedowninglaw.com/2023/07/13/ai-licensing-cant-balance-open-with-responsible/.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data?, 2023a.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023b.

The Free Software Foundation. What is free software?, 2024. URL https://web.archive.org/web/20230306010437/https://www.gnu.org/philosophy/free-sw.en.html. Last accessed on 2024-02-20.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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 Koh, Olga Saukh, Alexander 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, 2023.

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, 2024a.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024b.

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. *arXiv preprint arXiv:2111.09344*, 2021.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023a. URL https://zenodo.org/records/10256836.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023b.

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.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020b.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models, 2024.

Aidan Gomez. Introducing command r+: A scalable llm built for business. https://txt.cohere.com/command-r-plus-microsoft-azure/, 4 2024. URL: https://txt.cohere.com/command-r-plus-microsoft-azure/.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Suchin Gururangan, Mitchell Wortsman, Samir Yitzhak Gadre, Achal Dave, Maciej Kilian, Weijia Shi, Jean Mercat, Georgios Smyrnis, Gabriel Ilharco, Matt Jordan, Reinhard Heckel, Alex Dimakis, Ali Farhadi, Vaishaal Shankar, and Ludwig Schmidt. open_lm: a minimal but performative language modeling (lm) repository, 2023. URL https://github.com/mlfoundations/open_lm/. GitHub repository.

Laura Hanu and team Unitary. Detoxify, November 2020. URL https://github.com/unitaryai/detoxify.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. trlX: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8578–8595, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.530. URL https://aclanthology.org/2023.emnlp-main.530.

Luxi He, Mengzhou Xia, and Peter Henderson. What's in your" safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*, 2024.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *The Journal of Machine Learning Research*, 21(1): 10039–10081, 2020.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023.

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.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard, 2024.

Simon Hughes and Minseok Bae. Vectara Hallucination Leaderboard, November 2023. URL https://github.com/vectara/hallucination-leaderboard.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

The Open Source Initiative. The open source definition, February 2024. URL https://opensource.org/osd/.

Matthew Jagielski. A note on interpreting canary exposure. *arXiv preprint arXiv:2306.00133*, 2023.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12942–12950, 2023.

Yacine Jernite. Training data transparency in ai: Tools, trends, and policy recommendations. In *Hugging Face Blog*, 2023.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, et al. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2206–2222, 2022.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 7669–7673. IEEE, 2020.

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

Sayash Kapoor, Emily F. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica R. Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russel A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M Stewart, Gilles Vandewiele, and Arvind Narayanan. Reforms: Reporting standards for machine learning based science. *ArXiv*, abs/2308.07832, 2023. URL `https://arxiv.org/abs/2308.07832`.

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*, 2024.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. Glotlid: Language identification for low-resource languages, 2023.

Andrej Karpathy. nanogpt. GitHub Repo, 2023. URL `https://github.com/karpathy/nanoGPT`.

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. Golos: Russian dataset for speech research. *arXiv preprint arXiv:2106.10161*, 2021.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

Walter Klöpffer. Life cycle assessment: From the beginning to the current state. *Environmental Science and Pollution Research*, 4:223–228, 1997.

Kevin Klyman. Acceptable use policies for foundation models: Considerations for policymakers and developers. Stanford Center for Research on Foundation Models, apr 2024. URL `https://crfm.stanford.edu/2024/04/08/aups.html`.

Denis Kocetkov, Raymond Li, LI Jia, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, et al. The stack: 3 tb of permissively licensed source code. *Transactions on Machine Learning Research*, 2022.

Anton Korinek and Jai Vipra. Market concentration implications of foundation models: The invisible hand of chatgpt. 2023.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, et al. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text* (*W-NUT 2022*), pp. 215–221, 2022.

S Lakatos. A revealing picture: Ai-generated 'undressing'images move from niche pornography discussion forums to a scaled and monetized online business. Technical report, Technical report, Graphika, Dec 2023. URL https://public-assets. graphika . . . , 2023a.

Santiago Lakatos. A Revealing Picture: AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business. Technical report, December 2023b. URL `https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf`.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

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`.

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, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826, 2022.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024a.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024b.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. Learning from failure: Data capture in an australian aboriginal community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp. 4988–4998, 2022.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. 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. 8424–8445, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL `https://aclanthology.org/2022.acl-long.577`.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *Transactions on Machine Learning Research*, 2023a.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599, 2022b.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *arXiv preprint arXiv:2311.04287*, 2023b.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3197–3207, 2022.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021a.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. URL `https://aclanthology.org/2021.emnlp-demo.21`.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model. *arXiv preprint arXiv:2311.04219*, 2023a.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023b.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023c.

Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv preprint arXiv:2304.03271*, 2023d.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.

Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024b.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, 2020.

Shayne Longpre, Marcus Storm, and Rishi Shah. Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies. *MIT Science Policy Review*, 3(1):47–56, 2022.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023a.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023b.

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, et al. A safe harbor for ai evaluation and red teaming. *arXiv preprint arXiv:2403.04893*, 2024a.

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*, 2024b.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023a.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253), 2023b.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Marc Marone and Benjamin Van Durme. Data portraits: Recording foundation model training data. *arXiv preprint arXiv:2303.03919*, 2023.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging, 2022.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William A Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Daniel Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Y. Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. Dataperf: Benchmarks for data-centric AI development. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=LaFKTgrZMG`.

Daniel McDuff, Tim Korjakow, Scott Cambo, Jesse Josua Benjamin, Jenny Lee, Yacine Jernite, Carlos Muñoz Ferrandis, Aaron Gokaslan, Alek Tarkowski, Joseph Lindley, et al. On the standardization of behavioral use clauses and their adoption for responsible licensing of ai. *arXiv preprint arXiv:2402.05979*, 2024.

Kristina McElheran, J Frank Li, Erik Brynjolfsson, Zachary Kroff, Emin Dinlersoz, Lucia Foster, and Nikolas Zolas. Ai adoption in america: Who, what, and where. *Journal of Economics & Management Strategy*, 2024.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 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*, 2022.

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

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*, 2023a.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023b.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, pp. 2616–2620, 2017. doi: 10.21437/Interspeech.2017-950.

Arvind Narayanan and Sayash Kapoor. Ai safety is not a model property. `https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property`, March 2024. Accessed: YYYY-MM-DD.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*, 2023.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*, 2023.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. Better quality pre-training data and t5 models for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 158–168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.11. URL https://aclanthology.org/2023.emnlp-main.11.

OpenAccess-AI-Collective. Axolotl. GitHub Repo. URL https://github.com/OpenAccess-AI-Collective/axolotl.

OpenAI. Preparednessframework(beta), 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155.

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. IEEE, 2015.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2021.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*, 2023.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Guilherme Penedo, Alessandro Cappelli, Thomas Wolf, and Mario Sasko. Datatrove: large scale data processing, 2024. URL https://github.com/huggingface/datatrove.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Alexandra Sasha Luccioni, Yacine Jernite, and Anna Rogers. The roots search tool: Data transparency for llms. *arXiv preprint arXiv:2302.14035*, 2023.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pp. 2757–2761, 2020. doi: 10.21437/Interspeech.2020-2826.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21527–21536, 2024.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Inioluwa Deborah Raji and Roel Dobbe. Concrete problems in ai safety, revisited. *arXiv preprint arXiv:2401.10899*, 2023.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit, 2021.

Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*, 2023.

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, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 1–5. IEEE, 2023.

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

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

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. Codecarbon: estimate and track carbon emissions from machine learning computing. *Cited on*, pp. 20, 2021.

Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. apricot: Submodular selection for data summarization in python. *Journal of Machine Learning Research*, 21(161):1–6, 2020. URL `http://jmlr.org/papers/v21/19-467.html`.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022a.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022b.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020.

Riley Simmons-Edler, Ryan Badman, Shayne Longpre, and Kanaka Rajan. Ai-powered autonomous weapons risk geopolitical instability and threaten ai research. *arXiv preprint arXiv:2405.01859*, 2024.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9211, 2022a.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, 2022b.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Pete Walsh, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *Allen Institute for AI, Tech. Rep*, 2023.

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.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan

Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023b.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. Zipit! merging models from different tasks without training, 2024.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora* (*CMLC-7*). Leibniz-Institut für Deutsche Sprache, 2019.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Trustllm: Trustworthiness in large language models, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

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.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*, 2022.

David Thiel, Melissa Stroebel, and Rebecca Portnoff. Generative ML and CSAM: Implications and Mitigations. 2023a. doi: 10.25740/jv206yg3793. URL `https://purl.stanford.edu/jv206yg3793`.

David Thiel, Melissa Stroebel, and Rebecca Portnoff. Generative ml and csam: Implications and mitigations, 2023b.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pp. 2214–2218. Citeseer, 2012.

Together AI. Redpajama-data-v2: An open dataset with 30 trillion tokens for training large language models. Blog post on Together AI, Oct 2023. URL `https://www.together.ai/blog/redpajama-data-v2`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop* (*SLT*), pp. 652–658. IEEE, 2021.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the Third Workshop on Abusive Language Online*, pp. 80–93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3509. URL https://aclanthology.org/W19-3509.

Bertie Vidgen, Hannah Rose Kirk, Rebecca Qian, Nino Scherrer, Anand Kannappan, Scott A Hale, and Paul Röttger. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

Jai Vipra and Anton Korinek. Market concentration implications of foundation models: The invisible hand of chatgpt. *The Brookings Institution*, 2023. URL https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024.

Changhan Wang, Morgane Riviere, 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. *arXiv preprint arXiv:2101.00390*, 2021.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell (eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pp. 1–34, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.1. URL https://aclanthology.org/2023.conll-babylm.1.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. 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. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Herbert Woisetschläger, Alexander Isenko, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. A survey on efficient federated learning methods for foundation model training. *arXiv preprint arXiv:2401.04472*, 2024.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023a.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling, 2023b.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp. 3225–3245, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.181. URL https://aclanthology.org/2023.acl-long.181.

Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. *arXiv preprint arXiv:2401.12255*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models, 2023.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. Low-resource languages jailbreak gpt-4. In *Socially Responsible Language Modelling Research*, 2023.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-mix: a flexible and expandable family of evaluations for ai models. In *The Twelfth International Conference on Learning Representations*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 6182–6186. IEEE, 2022.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023a.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models with multiple choice questions, 2023b.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. (inthe) wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.

Jiahang Zhou, Yanyu Chen, Zicong Hong, Wuhui Chen, Yue Yu, Tao Zhang, Hui Wang, Chuanfu Zhang, and Zibin Zheng. Training and serving system of foundation models: A comprehensive survey. *arXiv preprint arXiv:2401.02643*, 2024.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024.

Piotr Żelasko, Daniel Povey, Jan "Yenda" Trmal, and Sanjeev Khudanpur. Lhotse: a speech data representation library for the modern deep learning ecosystem, 2021.

10 Cheatsheet Samples

In this portion of the appendix we have placed **samples** of the online cheatsheet, as examples of some of the content in each section.

Notations and Symbols The cheatsheet tables denote modalities and links with symbols, for brevity. Modalities are indicated for text T, vision V, and speech S. We also provide hyperlinks for ArXiv, for Hugging Face objects 🤗, for GitHub, and for webpages 🌐.

10.1 Data Sourcing

Modality	Name	Description	Links
		English Text	
T	**C4**	An English, cleaned version of Common Crawl's web crawl corpus (Raffel et al., 2020; Dodge et al., 2021).	
T	**Dolma**	An English-only pretraining corpus of 3 trillion tokens from a diverse mix of web content, academic publications, code, books, and encyclopedic materials (Soldaini et al., 2024). See the datasheet.	
T	**The Pile**	An 825GB English pretraining corpus that mixes portions of common crawl with 22 smaller, high-quality datasets combined together (Gao et al., 2020; Biderman et al., 2022).	
		Multilingual Text	
T	**ROOTS**	A massive multilingual pretraining corpus from BigScience, comprised of 1.6TB of text spanning 59 languages (Laurençon et al., 2022). It is a mix of OSCAR (Laippala et al., 2022) and the datasets found in the BigScience Catalogue (McMillan-Major et al., 2022).	
T	**MADLAD-400**	An general domain 3T token monolingual dataset based on CommonCrawl, spanning 419 languages (Kudugunta et al., 2023).	
T	**RedPajama v2**	A pretraining dataset of 30 trillion deduplicated tokens from 84 CommonCrawl dumps and 5 languages (Together AI, 2023).	
T	**CulturaX**	A pertaining dataset of 16T tokens, covering 167 languages, cleaned, deduplicated, and refined. Combines mC4 into 2020, with OSCAR project data up to 2023 (Nguyen et al., 2023).	
T	**OSCAR**	The Open Super-large Crawled Aggregated coRpus provides web-based multilingual datasets across 166 languages (Suárez et al., 2019; Laippala et al., 2022).	
T	**WURA**	A manually audited multilingual pre-training corpus (document-level dataset) for 16 African languages and four high-resource languages widely spoken in Africa (English, French, Arabic and Portuguese) (Oladipo et al., 2023).	

Modality	Name	Description	Links
		Vision Pretraining Corpora	
T V	**LAION-5B**	A collection of over 5B image-text pairs collected from Common Crawl, optionally English-filtered (Schuhmann et al., 2022b).	
T V	**DataComp & CommonPool**	A large pool of 13B image-text pairs from CommonCrawl and a curated 1B subset (Gadre et al., 2023).	
T V	**MMC4**	Interleaved image-text data from Common Crawl (570M images, 43B tokens) (Zhu et al., 2023).	
T V	**OBELICS**	Interleaved image-text data from Common Crawl (353 M images, 115B tokens) (Laurençon et al., 2023).	
		Speech Pretraining Corpora	
S	**Common Voice**	28k hours of crowd-sourced read speech from 100+ languages.	
S	**GigaSpeech**	40k hours (10k transcribed) multi-domain English speech corpus (Chen et al., 2021).	
S	**Golos**	1,240 hours of crowd-sourced Russian speech (Karpov et al., 2021).	
S	**IndicSUPERB**	1,684 hour crowd-sourced corpus of 12 Indian languages (Javed et al., 2023).	
S	**Libri-Light**	LibriVox English audiobooks (60k hours) (Kahn et al., 2020).	
S	**The People's Speech**	30k hour conversational English dataset (Galvez et al., 2021).	
S	**VoxPopuli**	400k hours of unlabelled speech from 23 languages of the European parliament (Wang et al., 2021).	
S	**WenetSpeech**	22.4k hour multi-domain corpus of Mandarin (Zhang et al., 2022).	

Modality	Name	Description	Links
		Specialized Text Pretraining Corpora	
T	**Open License Corpus**	The Open License Corpus is a 228B token corpus of permissively-licensed, English text data for pretraining (Min et al., 2023).	
T	**Pile of Law**	An open-source, English dataset with 256GB of legal and administrative data, covering court opinions, contracts, administrative rules, and legislative records (Henderson et al., 2022).	
T	**The Stack**	A 6TB permissively-licensed corpus from active GitHub repositories (358 programming languages) (Kocetkov et al., 2022).	
T	**The Proof Pile 2**	The Proof-Pile-2 is a 55 billion token dataset of mathematical and scientific documents (Azerbayev et al., 2023).	
T	**peS2o**	A collection of 40M creative open-access academic papers, cleaned, filtered, and formatted for pre-training of language models, originally derived from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020).	

Modality	Name	Description	Links
		Finetuning Data Catalogs	
T S	**Arabic NLP Data Catalogue**	A catalogue of hundreds of Arabic text and speech finetuning datasets, regularly updated.	
T	**Data Provenance Collection**	An explorer tool for selecting popular finetuning, instruction, and alignment training datasets from Hugging Face, based on data provenance and characteristics criteria Longpre et al. (2023b).	
V	**ImageNet**	An image classification dataset with 1.3M samples and 1000 classes (Russakovsky et al., 2015).	
T S	**Masakhane NLP**	A repository of African language text and speech resources, including datasets.	
T V	**MS COCO**	Object detection, segemndation, captioning and retrieval dataset (Lin et al., 2014).	
S	**OpenSLR**	A collection of user-contributed datasets for various speech processing tasks.	
T S	**SEACrowd**	A repository of hundreds of South East Asian language datasets.	
S	**VoxLingua107**	Spoken language identification dataset created from YouTube videos, categroized by search phrases (Valk & Alumäe, 2021).	
T S	**Zenodo AfricaNLP Community**	An online catalogue that provides African language resources (data and models) in both texts and speech.	

10.2 Data Preparation

Modality	Name	Description	Links
		Pretraining Data Search & Analysis	
T	**AI2 C4 Search Tool**	A search tool that lets users to execute full-text queries to search Google's C4 Dataset.	
T V	**LAION search**	Nearest neighbor search based on CLIP embeddings.	
T	**ROOTS Search Tool**	A tool, based on a BM25 index, to search over text for each language or group of languages included in the ROOTS pretraining dataset (Piktus et al., 2023).	
T	**WIMBD**	A dataset analysis tool to count, search, and compare attributes across several massive pretraining corpora at scale, including C4, The Pile, and RedPajama Elazar et al. (2023b).	
		Finetuning Data Search & Analysis	
T	**Data Provenance Explorer**	An explorer tool for selecting, filtering, and visualizing popular finetuning, instruction, and alignment training datasets from Hugging Face, based on their metadata such as source, license, languages, tasks, topics, among other properties Longpre et al. (2023b).	
V	**Know Your Data**	A tool for exploring over 70 vision datasets.	
S	**NVIDIA Speech Data Explorer**	Tool for exploring speech data.	

Modality	Name	Description	Links
		Dataset Cleaning & Filtering	
T	**Dolma's Toolkit**	A Python framework for defining Taggers that identify non-language text, language ID, PII, toxic text, and "quality" text. Includes reimplementation of heuristics used by Gopher and C4 for non-natural language (Soldaini et al., 2023).	
T V	**DataComp pre-filtering**	NSFW detection, dedup with eval datasets (Gadre et al., 2023).	
T	**Lilac**	A python package for better understanding your data. Includes keyword and semantic search, as well as detection for PII, duplicates, and language.	
T	**Roots data cleaning pipeline**	A pipeline for processing and improving quality of crowdsourced datasets (Laurençon et al., 2022).	
		Language Identification	
T	**Langdetect**	A tool to predict the language of text, used to filter out/in data from the desired languages.	
T	**fastText language classifier**	A tool for classifying the language of text (Grave et al., 2018).	
T	**FUN-LangID**	Frequently Used N-grams Language ID model, a character 4-gram model trained to recognize up to 1633 languages.	
T	**OpenLID**	A model (and data used to train the model) for identifying 200+ languages (Burchell et al., 2023).	
T	**GlotLID**	A model for identifying languages, with support for more than 1600 languages (Kargaran et al., 2023).	
S	**SpeechBrain's Spoken language ID model**	Pre-trained spoken language identification model trained on VoxLingua107, dataset of audio sourced from YouTube for 107 languages (Ravanelli et al., 2021).	

Modality	Name	Description	Links
		Dataset Deduplication	
T V S	**Apricot**	Apricot implements submodular optimization for summarizing massive datasets into minimally redundant subsets, useful for visualizing or deduplicating datasets (Schreiber et al., 2020).	
V	**DataComp**	Data to deduplicate vision datasets for the DataComp challenge (Gadre et al., 2023).	
T	**Dolma's Toolkit**	Dolma uses a Bloom Filter implemented in Rust (Soldaini et al., 2023).	
T	**Google Text Deduplication**	A repository to deduplicate language datasets (Lee et al., 2022a).	
T	**Pile Deduplication**	A set of tools for MinHashLSH deduplication (Gao et al., 2020).	

Modality	Name	Description	Links
		Dataset Decontamination	
T	**BigBench Canaries**	BigBench's "Training on the Test Set" Task provies guidance on using canaries to check if an evaluation set was trained on.	
T	**Carper AI Decontamination Tool**	A repository, heavily based by the BigCode repository, to decontaminate evaluation sets from a text training set.	
T	**Data Portraits**	A tool to test if a model has seen certain data, e.g. in the The Pile or The Stack (Marone & Van Durme, 2023).	
T	**Detect Data (Min-K Prob)**	A codebase that implements "Min-K% Prob", a method to detect if a language model was pretrained on some text (Shi et al., 2023).	

Modality	Name	Description	Links
		Data Auditing Tools	
T V	**HaveIBeenTrained**	A combination search tool / opt out tool for LAION	
		Data Auditing Case Studies	
T	**A Datasheet for BookCorpus**	A third party datasheet for BookCorpus (Bandy & Vincent, 2021).	
T	**Data Provenance Initiative**	A large scale audit of 2000+ popular datasets in AI (Longpre et al., 2023b).	
T	**Datasheet for the Pile**	A datasheet for the Pile (Biderman et al., 2022).	
T V	**Into the LAIONs Den**	Auditing hateful content in text-to-vision datasets (Birhane et al., 2023b).	
T V	**Multimodal dataset audit**	Auditing vision datasets for highly sensitive content, including misogyny, pornography and malignant stereotypes (Birhane et al., 2021).	
T V	**On Hate Scaling Laws**	Auditing text and vision datasets for systemic biases and hate (Birhane et al., 2023a).	
T	**Quality at a Glance**	An audit of allegedly multilingual parallel text corpora (Kreutzer et al., 2022).	

10.3 Data Documentation and Release

Modality	Name	Description	Links
		Data Documentation	
T V S	**Data Cards Playbook**	A tool to create a Data Card that thoroughly documents a new dataset (Pushkarna et al., 2022).	
T V S	**Data Provenance Attribution Card**	A repository to select datasets and generate a summary. It can also generate a bibtex to attribute all developers of the datasets (Longpre et al., 2023b).	
T V S	**Datasheets for Datasets**	A datasheet to thoroughly document a new dataset (Gebru et al., 2021).	
T V S	**Datasheets for Digital Cultural Heritage Datasets**	A datasheet specifically designed for digital cultural heritage datasets and their considerations (Alkemade et al., 2023).	

Modality	Name	Description	Links
		Data Governance	
T V S	**Data Governance for BLOOM**	A paper detailing the data governance decisions undertaken during BigScience's BLOOM project. Jernite et al. (2022).	
T V S	**Data Trusts for Training Data**	A paper that argues for the creation of a public data trust for collective input into the creation of AI systems and analyzes the feasibility of such a data trust. Chan et al. (2023).	
T V	**HaveIBeenTrained**	A combination search tool / opt out tool for LAION.	

10.4 Model Training

Modality	Name	Description	Links
		Pretraining Repositories	
T	**GPT-NeoX**	A library for training large language models, built off Megatron-DeepSpeed and Megatron-LM with an easier user interface. Used at massive scale on a variety of clusters and hardware setups (Andonian et al., 2021).	
T	**Megatron-DeepSpeed**	A library for training large language models, built off of Megatron-LM but extended by Microsoft to support features of their DeepSpeed library (Smith et al., 2022b).	
T	**OpenLM**	OpenLM is a minimal language modeling repository, aimed to facilitate research on medium sized LMs. They have verified the performance of OpenLM up to 7B parameters and 256 GPUs. They only depend only on PyTorch, XFormers, or Triton (Gururangan et al., 2023).	
T V	**Kosmos-2**	For training multimodal models with CLIP backbones (Peng et al., 2023).	
T V	**OpenCLIP**	Supports training and inference for over 100 CLIP models (Ilharco et al., 2021).	
V	**Pytorch Image Models (timm)**	Hub for models, scripts and pre-trained weights for image classification models.	
S	**Lhotse**	Python library for handling speech data in machine learning projects.	
S	**Stable Audio Tools**	A codebase for distributed training of generative audio models.	

Modality	Name	Description	Links
		Finetuning Repositories	
T	**Axolotl**	A repository for chat- or instruction-tuning language models, including through full fine-tuning, LoRA, QLoRA, and GPTQ.	
T	**Levanter**	A framework for training large language models (LLMs) and other foundation models that strives for legibility, scalability, and reproducibility.	
T V	**LLaMA-Adapter**	Fine-tuned LLMs on multimodal data using adapters (Gao et al., 2023b).	
T V	**LLaVA**	Fine-tuned LLMs on multimodal data using a projection layer (Liu et al., 2023a).	
T V	**Otter**	Multimodal models with Flamingo architecture (Li et al., 2023a).	
T V S	**peft**	A library for doing parameter efficient finetuning.	
T V S	**trlX**	A library for doing RLHF at scale (Havrilla et al., 2023).	

Modality	Name	Description	Links
		Efficiency and Resource Allocation	
T V S	**Cerebras Model Lab**	A calculator to apply compute-optimal scaling laws for a given budget, including factoring expected total inference usage.	
T V S	**QLoRa**	An efficient finetuning approach that reduces memory usage while training (Dettmers et al., 2023).	
T	**Scaling Data-Constrained Language Models**	Demonstrates an optimal allocation of compute when dataset size is bounded (Muennighoff et al., 2023b).	
T	**Training Compute-Optimal Language Models**	Proposes an optimal allocation of computational budget between model and dataset size, and shows experimental design for fitting scaling laws for compute allocation in a new setting (Hoffmann et al., 2022).	

Modality	Name	Description	Links
T V S	**The EleutherAI Model Training Cookbook**	A set of resources on how to train large scale AI systems (Anthony et al., 2024).	
T V S	**Machine Learning Engineering Online Book**	An "online textbook" and resource collection on ML engineering at scale, ranging from debugging distributed systems, parallelism strategies, effective use of large HPC clusters, and chronicles of past large-scale training runs with lessons learned.	
T	**nanoGPT**	A minimal, stripped-down training codebase for teaching purposes and easily-hackable yet performant small-scale training.	
T	**Transformer Inference Arithmetic**	A blog post on the inference costs of transformer-based LMs. Useful for providing more insight into deep learning accelerators and inference-relevant decisions to make when training a model.	
T	**Transformer Math 101**	An introductory blog post on training costs of LLMs, going over useful formulas and considerations from a high to low level.	

10.5 Environmental Impact

Modality	Name	Description	Links
		Estimating Environmental Impact	
T	**Estimating the Carbon Footprint of BLOOM**	A comprehensive account of the broader environmental impact of the BLOOM language model (Luccioni et al., 2023b).	
T V S	**Azure Emissions Impact Dashboard**	Monitoring the environmental impact of training machine learning models on Azure.	
T V S	**Carbontracker**	carbontracker is a tool for tracking and predicting the energy consumption and carbon footprint of training deep learning models (Anthony et al., 2020).	
T V S	**CodeCarbon**	Estimate and track carbon emissions from your computer, quantify and analyze their impact (Schmidt et al., 2021).	
T V S	**Experiment Impact Tracker**	The experiment-impact-tracker is meant to be a drop-in method to track energy usage, carbon emissions, and compute utilization of your machine learning workload Henderson et al. (2020).	
T V S	**Google Cloud Carbon Footprint Measurement**	Tracking the emissions of using Google's cloud compute resources.	
T V S	**Making AI Less "Thirsty"**	Uncovering and Addressing the Secret Water Footprint of AI Models, and estimating water usage for training and deploying LLMs.	
T V S	**ML CO2 Impact**	A tool for estimating carbon impacts of ML training (Lacoste et al., 2019).	

Modality	Name	Description	Links
		Effective Use of Resources	
T	**Scaling Laws for Neural Language Models**	Provide scaling laws to determine the optimal allocation of a fixed compute budget.	
T	**Training Compute-Optimal Large Language Models**	Provides details on the optimal model size and number of tokens for training a transformer-based language model in a given computational budget.	

10.6 Environmental Impact

Modality	Name	Description	Links
		Common Benchmarks for Text Capability Evaluation	
T	**BigBench**	A collaborative benchmark of 100s of tasks, probing LLMs on a wide array of unique capabilities (Srivastava et al., 2023a).	
T	**BigBench Hard**	A challenging subset of 23 BigBench tasks where 2022 models did not outperform annotator performance (Suzgun et al., 2022).	
T	**HELM classic**	A comprehensive suite of scenarios and metrics aimed at holistically evaluating models (including for capabilities) with comparisons to well known models (Liang et al., 2022).	
T	**HELM lite**	A lightweight subset of capability-centric benchmarks within HELM. Compares prominent open and closed models.	
T	**LM Evaluation Harness**	Orchestration framework for standardizing LM prompted evaluation, supporting hundreds of subtasks.	
T	**MMLU**	Evaluation of LM capabilities on multiple-choice college exam questions (Hendrycks et al., 2020).	
T	**MTEB**	The Massive Text Embedding Benchmark measures the quality of embeddings across 58 datasets and 112 languages for tasks related to retrieval, classification, clustering or semantic similarity.	
T	**SIB-200**	A large-scale open-sourced benchmark dataset for topic classification in 200 languages and dialects (Adelani et al., 2023).	
		Common Benchmarks for Code Capability Evaluation	
T	**BigCode Evaluation Harness**	A framework for the evaluation of code generation models, compiling many evaluation sets.	
T	**HumanEvalPack**	A code evaluation benchmark across 6 languages and 3 tasks, extending OpenAI's HumanEval (Muennighoff et al., 2023a).	
T	**SWE Bench**	SWE-bench is a benchmark for evaluating large language models on real world software issues collected from GitHub. Given a codebase and an issue, a language model is tasked with generating a patch that resolves the described problem (Jimenez et al., 2023).	

Modality	Name	Description	Links
		Common Benchmarks for Multimodal Capability Evaluation	
T V	**CLIP benchmark**	Image classification, retrieval and captioning.	
T V	**DataComp eval suite**	38 image classification and retrieval downstream tasks (Gadre et al., 2023).	
T V	**HEIM**	A large suite of text-to-image evaluations. Useful for thorough capability analysis of these model types (Lee et al., 2023b).	
S	**The Edinburgh International Accents of English Corpus**	Benchmark dataset of diverse English varieties for evaluating automatic speech recognition models (typically trained and tested only on US English) (Sanabria et al., 2023).	
T V	**MMBench**	A joint vision and text benchmark evaluating dozens of capabilities, using curated datasets and ChatGPT in the loop (Liu et al., 2023b).	
T V	**MME**	An evaluation benchmark for multimodal large language models with 14 manually curated subtasks, to avoid data leakage (Fu et al., 2023).	
T V	**MMMU**	A benchmark to evaluate joint text and vision models on 11k examples spanning 30 college-level subject domains (Yue et al., 2023).	
T V	**OpenFlamingo eval suite**	VQA, captioning, classification evaluation suite (Awadalla et al., 2023).	
		Common Leaderboards for Capability Evaluation	
T V S	**Hugging Face Leaderboards**	A set of popular model leaderboards on Hugging Face for ranking on generic metrics.	
T	**LMSys Chatbot Arena**	A leaderboard of models based on Elo ratings where humans or models select their preferred response between two anonymous models. Chatbot Arena, MT-Bench, and 5-shot MMLU are used as benchmarks. This resource provides a general purpose, and GPT-4 biased perspective into model capabilities (Zheng et al., 2023).	
S	**OpenASR Leaderboard**	An automatic leaderboard ranking and evaluating speech recognition models on common benchmarks.	

Modality	Name	Description	Links
		Taxonomies of Risk and Harm	
T V S	**Ethical and social risks of harm from language models**	A two-level taxonomy of LLM risks, comprising both classification groups and associated harms (Weidinger et al., 2021). The classification groups are: (1) Discrimination, Exclusion and Toxicity, (2) Information Hazards, (3) Misinformation Harms, (4) Malicious Uses, (5) Human-Computer Interaction Harms, and (6) Automation, access, and environmental harms.	
T V S	**Taxonomy of Risks posed by Language Models**	A taxonomy of language model risks (Weidinger et al., 2022). The classification groups are: (1) Discrimination, Hate speech and Exclusion, (2) Information Hazards, (3) Misinformation Harms, (4) Malicious Uses, (5) Human-Computer Interaction Harms, and (6) Environmental and Socioeconomic harms. For each risk area, the authors describe relevant evidence, causal mechanisms, and risk mitigation approaches.	
T V S	**Sociotechnical Safety Evaluation of Generative AI Systems**	A two-level taxonomy of AI risks, comprising both classification groups and associated harms (Weidinger et al., 2023). The classification groups are: (1) Representation and Toxicity Harms, (2) Misinformation Harms, (3) Information & Society Harms, (4) Malicious Use, (5) Human Autonomy & Integrity Harms, and (6) Socioeconomic & Environmental Harms.	
T V S	**Sociotechnical Harms of Algorithmic Systems**	A taxonomy of AI harms with detailed subcategories, focused on mitigating harm (Shelby et al., 2023). The harm categories are: (1) Representational harms, (2) Allocative harms, (3) Quality of Service harms, (4) Interpersonal harms, and (5) Social system harms.	
T V S	**Assessing Language Model Deployment with Risk Cards**	A framework for structured assessment and documentation of risks associated with applications of language models (Derczynski et al., 2023). Each RiskCard makes clear the routes for the risk to manifest harm, their placement in harm taxonomies, and example prompt-output pairs. 70+ risks are identified, based on a literature survey.	
T V S	**An Overview of Catastrophic AI Risks**	A taxonomy of 4 catastrophic AI risks, with associated subcategories (Hendrycks et al., 2023): (1) Malicious use,(2) AI race,(3) Organizational risks, (4) Rogue AIs (Proxy gaming, Goal drift, Power-seeking, Deception).	
T V S	**OpenAI Preparedness Framework**	Description of 4 catastrophic AI risks (OpenAI, 2023): (1) Cybersecurity, (2) Chemical, Biological, Nuclear and Radiological (CBRN) threats, (3) Persuasion, and (4) Model autonomy. The paper also highlights the risk of "unknown unknowns".	
T V S	**Model evaluation for extreme risks**	A framework of 9 dangerous capabilities of AI models (Shevlane et al., 2023): (1) Cyber-offense, (2) Deception, (3) Persuasion & manipulation, (4) Political strategy, (5) Weapons acquisition, (6) Long-horizon planning, (7) AI development, (8) Situational awareness, (9) Self-proliferation.	

Modality	Name	Description	Links
		Risk Evaluation Overviews	
T	**SafetyPrompts website**	Open repository of datasets for LLM evaluation and mitigation.	
T V S	**Safety evaluation repository**	A repository of 200+ safety evaluations, across modalities and harms. Useful for delving deeper into the array of risks.	
		Bias & Toxicity Evaluations	
T	**Bias Benchmark for QA (BBQ)**	A dataset of question-sets constructed by the authors that highlight attested social biases against people belonging to protected classes along nine different social dimensions relevant for U.S. English-speaking contexts (Parrish et al., 2021).	
T V	**Crossmodal-3600**	Image captioning evaluation with geographically diverse images in 36 languages (Thapliyal et al., 2022).	
T	**HolisticBias**	A bias and toxicity benchmark using templated sentences, covering nearly 600 descriptor terms across 13 different demographic axes, for a total of 450k examples (Smith et al., 2022a).	
T	**RealToxicityPrompts**	100k web sentence snippets for researchers to assess the risk of neural toxic degeneration in models (Gehman et al., 2020b).	
V	**StableBias**	Bias testing benchmark for Image to Text models, based on gender-occupation associations (Luccioni et al., 2023a).	
		Factuality Evaluations	
T	**FactualityPrompt**	A benchmark to measure factuality in language models (Lee et al., 2022b).	
T	**Hallucinations**	Public LLM leaderboard computed using Vectara's Hallucination Evaluation Model. It evaluates LLM hallucinations when summarizing a document (Hughes & Bae, 2023).	
		Information & Safety Hazard Evaluations	
T	**Purple Llama CyberSecEval**	A benchmark for coding assistants, measuring their propensity to generate insecure code and level of compliance when asked to assist in cyberattacks (Bhatt et al., 2023).	
T	**Purple Llama Guard**	A tool to identify and protect against malicious inputs to LLMs (Inan et al., 2023).	
T V S	**OpenAI content moderation filter**	A moderation filter and released dataset. The endpoint has 5 primary categories (Sexual, Hateful, Violent, Self-harm, Harassment) with sub-categories.	
T	**SimpleSafetyTests**	Small probe set (100 English text prompts) covering severe harms: child abuse, suicide, self-harm and eating disorders, scams and fraud, illegal items, and physical harm (Vidgen et al., 2023).	

10.7 Model Release & Monitoring

Modality	Name	Description	Links
		Model Documentation	
T V S	**Model Cards**	A standard for reporting and documenting machine learning models, for promoting and easing transparent and open model development or reporting Mitchell et al. (2019).	
T V S	**Model Card Resources**	A release of several resources surrounding model cards, including templates and tools for easy documentation creation, and how these are frequently used in practice.	
T V S	**Ecosystem Cards**	Ecosystem Graphs centralize information about models and their impact in the broader ecosystem Bommasani et al. (2023b).	
T V S	**Foundation Model Transparency Index**	An index to measure the transparency of a foundation model with respect to its inputs, development, and downstream uses or policies Bommasani et al. (2023a).	

Modality	Name	Description	Links
		Code Reproducibility	
T V S	**Anaconda**	An environment and dependency management tool.	
T V S	**Colab Notebooks**	A tool to execute and share reproducible code snippets.	
T V S	**Docker**	An environment and dependency management tool.	
T V S	**Jupyter Notebooks**	A tool to execute and share reproducible code snippets.	
T	**LM Evaluation Harness**	Orchestration framework for standardizing LM prompted evaluation, supporting hundreds of subtasks (Gao et al., 2023a).	
T V S	**Semver**	A widely used protcol for versioning to software, to ensure easy reproducibility.	

Modality	Name	Description	Links
		General Guidance	
T V S	**Behavioral Use Licensing for Responsible AI**	A paper that provides a theoretical framework for licenses intended for open models with use restrictions (Contractor et al., 2022).	
T V S	**Legal Playbook For Natural Language Processing Researchers**	This playbook is a legal research resource for various activities related to data gathering, data governance, and disposition of an AI model available as a public resource.	
T V S	**The Turning Way, Licensing**	A guide to reproducible research and licensing.	

Modality	Name	Description	Links
		License Selection Guides	
T V S	**Choose an open source license**	A guide for choosing among open source licenses that includes general selection criteria and explanations for software licenses.	
T V S	**Creative Commons License Chooser**	A guide for choosing among Creative Commons licenses with an explanation of how they function.	
T V S	**Primer on AI2 ImpACT Licenses**	A post by AI2 describing when and why an organization should use a specific ImpACT license.	
T V S	**Primer on RAIL Licenses**	A post by RAIL describing when and why an organization should use a specific RAIL license.	
T V S	**License Clause Comparison**	A comparison of behavioral use clauses included in foundation model licenses (McDuff et al., 2024)	
		Licenses	
T V S	**Apache 2.0 License**	The most common open-source license for model weights.	
T V S	**AI2 ImpACT-LR License**	License for low risk AI artifacts (data and models) that allows for distribution of the artifact and its derivatives. Use restrictions include weapons development and military surveillance.	
T V S	**AI2 ImpACT-MR License**	License for medium risk AI artifacts (data and models) that does not allows for distribution of the artifact but does allow for distribution of its derivatives. Use restrictions include weapons development and military surveillance.	

Modality	Name	Description	Links
T V S	**AI Vulnerability Database**	An open-source, extensible knowledge base of AI failures.	
T V S	**Llama 2 Responsible Use Guide**	Guidance for downstream developers on how to responsibly build with Llama 2. Includes details on how to report issues and instructions related to red-teaming and RLHF.	
T V S	**BigScience Ethical Charter**	Outlines BigScience's core values and how they promote them, which in turn guides use restrictions.	
T V S	**Model Monitoring in Practice Tutorial**	A tutorial given at FAccT and other venues describing how and why to monitor ML models. Includes a presentation on using transformer models to monitor for error detection.	
T V S	**Model Gating from Hugging Face**	A resource describing how to require user credentials for model access, which may be appropriate for models trained for topics such as hate speech.	