The Foundation Model Development Cheatsheet

Shayne Longpre[1*] **Stella Biderman**[2*]
Alon Albalak[3] **Gabriel Ilharco**[4] **Sayash Kapoor**[5]
Kevin Klyman[6,7] **Kyle Lo**[8] **Maribeth Rauh**[9]
Nay San[6] **Hailey Schoelkopf**[2] **Aviya Skowron**[2]
Bertie Vidgen[10] **Laura Weidinger**[9] **Arvind Narayanan**[5]
Victor Sanh[11] **David Adelani**[12] **Percy Liang**[6]
Rishi Bommasani[6] **Peter Henderson**[5] **Sasha Luccioni**[11]
Yacine Jernite[11*] **Luca Soldaini**[8*]

[1]MIT [2]EleutherAI [3]UCSB [4]UW [5]Princeton University
[6]Stanford University [7]Harvard University [8]Allen Institute for AI
[9]Google DeepMind [10]ML Commons [11]HuggingFace
[12]University College London, Masakhane

fmcheatsheet.org

A Guide: Resources for Best Development Practices

Why? This cheatsheet serves as a succinct guide, prepared *by* foundation model developers *for* foundation model developers. As the field of AI foundation model development rapidly expands, welcoming new contributors, scientists, and applications, we hope to lower the barrier for new community members to become familiar with the variety of resources, tools, and findings. The focus of this cheatsheet is not only, or even primarily, to support building, but to inculcate good practices, awareness of limitations, and general responsible habits as community norms. While it is certainly not comprehensive, we have selected a sample of resources that we have found useful and would recommend for consideration by others. We hope it will serve as a general guide to promote responsible development practices, as well as building new models and infrastructure in our field. This document provides contextualized information and a static sample of the cheatsheet—the fully documented, live cheatsheet is available at `fmcheatsheet.org`.

What? There are many exceedingly popular tools to build, distribute and deploy foundation models. But there are also many outstanding resources that receive less attention or adoption, in part because of developers' haste to accelerate, deploy, and monetize. We hope to bring wider attention to these core resources that support *informed* data selection, processing, and understanding (§§ 1 and 2), *precise and limitation-aware* artifact documentation (Section 3), *efficient* model training (Section 4), *advance awareness* of the environmental impact from training (Section 5), *careful* model evaluation of capabilities, risks, and claims (Section 6), as well as *responsible* model release and deployment practices (Section 7).

In each section we introduce considerations for that phase of development. We suggest developers building models from scratch carefully consider their data sources and curation processes to avoid unintended risk (e.g. privacy, copyright, inappropriate content), marginalization (e.g. by accidental filtering), or unexpected model behaviors (e.g. text distribution has unintended quirks). Data processing should consider how to deduplicate, filter, decontaminate, and mix different data sources towards the intended distribution and characteristics. We recommend this process is carefully documented for reproducibility and understanding. If new data is released, setting its governance standards early will avoid misuse later, and adding structure to documentation will allow for its properties to be easily preserved, understood, and respected in downstream data mixes or compositions. Model training can be financially and environmentally expensive. Resources for estimating environmental impact can break down these costs and simplify the considerations. Newly trained models should be carefully evaluated for their intended uses, as

*Equal contribution. Correspondence: slongpre@mit.edu.
Note that all authors co-led at least one section or modality of the cheatsheet, and as such the core contributors are listed alphabetically by last name (see Appendix A).

well as foreseeable misuses or harms. We suggest resources to taxonomize and contextualize evaluations, without over-claiming or misunderstanding the limitations of the reported numbers. Lastly, we suggest how developers can make an informed selection of licenses, and release mechanisms, to mitigate potential misuses.

Criteria for Inclusion. 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 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 1.2 we try to include more Finetuning Data Catalogs for lower resource languages, that often receive less attention than HuggingFace's Dataset library. Rather than sharing primarily academic literature as in a traditional survey work, we focus on tools, such as data catalogs, search/analysis tools, evaluation repositories, and, selectively, literature that summarizes, surveys or guides important development decisions. As with any survey, these principles for inclusion are incomplete and subjective, but we hope to remedy this with the open call for community contributions.

How to contribute? We intend for the web version of the cheatsheet to be a crowdsourced, interactive, living document, with search and filter tools. To contribute to this resource, follow instructions given in the website. Contributions should be scoped to resources, such as tools, artifacts, or helpful context papers, that directly inform responsible foundation model development. Significant contributions will be recognized in the web search tool, and in follow up write-ups to this guide. Contributed content will be reviewed by authors, to assess that the resources fit the aspirational criteria for inclusion outlined above.

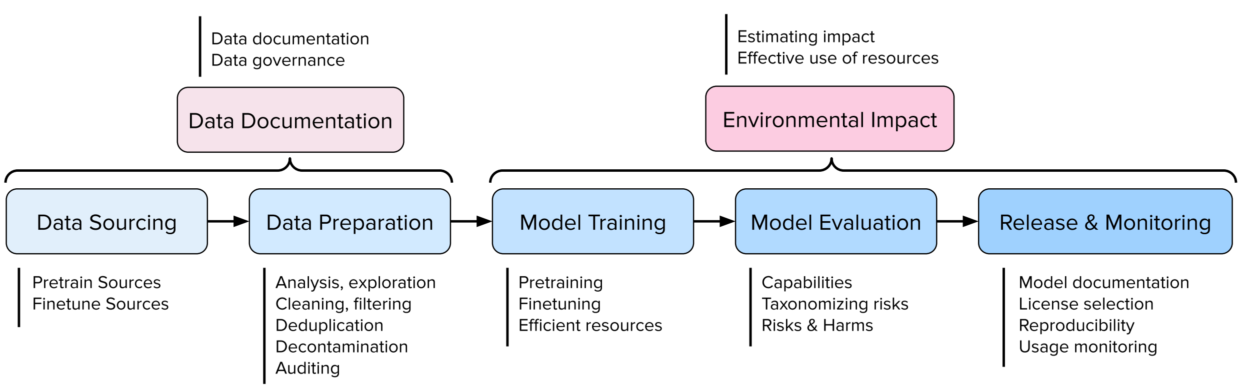
Scope & Limitations. We've compiled resources, tools, and papers that have helped guide our own intuitions around model development, and which we believe will be especially helpful to nascent (and sometimes even experienced) developers in the field. However, this guide is far from exhaustive—and here's what to consider when using it:

- We scope these resources to newer foundation model developers, usually releasing models to the community. Larger organizations, with commercial services, have even broader considerations for responsible development and release.
- Foundation model development is a rapidly evolving science. **This document is only a sample, dated to January 2024.** The full cheatsheet is more comprehensive and open for on-going public contributions: `fmcheatsheet.org`.
- We've scoped our data modalities only to **text, vision, and speech**. We support multilingual resources, but acknowledge this is only a starting point.
- A cheatsheet **cannot be comprehensive**. We prioritize resources we have found helpful, and rely heavily on survey papers and repositories to point out the many other awesome works which deserve consideration, especially for developers who plan to dive deeper into a topic.
- **We cannot take responsibility for these resources—onus is on the reader to assess their viability, 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 readers judgement.

Notations and Symbols This cheatsheet provides tables for several topics, 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.

Who are we? An organic collective of volunteers have contributed to this cheatsheet, spanning developers of several notable datasets (MasakhaNER (Adelani et al., 2021), the Pile (Gao et al., 2020), ROOTS (Laurençon et al., 2022), Dolma (Soldaini et al., 2023), the Flan Collection (Longpre et al., 2023a), the Data Provenance Collection (Longpre et al., 2023b)), models (OpenFlamingo (Awadalla et al., 2023), Pythia (Biderman et al., 2023), Flan-PaLM (Chung et al., 2022), RWKV (Peng et al., 2023a)), and benchmarks (LM Eval Harness (Gao et al., 2023a), HELM (Liang et al., 2022)) in the community. Most importantly, contributors (Appendix A) have carefully assembled the tools and resources that have enabled them to build this infrastructure responsibly with an eye on social and scientific impact.

Contents



Data documentation
Data governance
Data Documentation
Estimating impact
Effective use of resources
Environmental Impact
Data Sourcing
Data Preparation
Model Training
Model Evaluation
Release & Monitoring
Pretrain Sources
Finetune Sources
Analysis, exploration
Cleaning, filtering
Deduplication
Decontamination
Auditing
Pretraining
Finetuning
Efficient resources
Capabilities
Taxonomizing risks
Risks & Harms
Model documentation
License selection
Reproducibility
Usage monitoring

1 Data Sources

Data Sourcing Best Practices

- Pretraining data provides the fundamental ingredient to foundation models—including their capabilities and flaws. Finetuning data hones particular abilities of a model, or in the case of instruction finetuning or alignment training, improves the models general usability and helpfulness while reducing 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 hone some capabilities or impair others. Use catalogs to support an informed selection, and prefer well-documented to under-documented datasets.
- The most appropriate datasets may not exist for a given set of tasks. Be aware of the limitations of choosing from what is available.

1.1 Pretraining Data Sources

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).	

Pretraining data consists of thousands, or even millions, of individual documents, often web scraped. As a result, their contents are often superficially documented or understood. Model knowledge and behavior will likely reflect a compression of this information and its communication qualities. Consequently, its important to carefully select the data composition. This decision should reflect choices in the language coverage, the mix of sources, and preprocessing decisions. We highlight a few of the most popular pretraining corpora which have accumulated deeper documentation or analyses.

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., 2022).	
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).	

Specialized pretraining sources can mitigate model risks and cater them to certain capabilities. The examples we provide focus on coding data, mathematical or scientific data, legal data, or restrict to data without copyright concerns.

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).	

1.2 Finetuning Data Catalogs

Finetuning data is 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 use a variety of data annotations and loss objectives for traditional supervised finetuning, DPO (Rafailov et al., 2023) or reinforcement learning objectives (Ouyang et al., 2022). As a result of this variety, we recommend data catalogs, with attached documentation, to help an informed selection. The largest catalog is HuggingFace Datasets (Lhoest et al., 2021), though cross-reference its metadata with academic papers and repositories, as its crowdsourced documentation can be sparse or incorrect.

Aside from HuggingFace Datasets, we point to some lesser known resources that catalog more specialized finetuning data sources. The breadth of available finetuning data is expansive, so we focus on catalogs rather than individual datasets, and particularly those that provide strong documentation or more specialized sources.

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 and Alumäe, 2021).	
T S	**Zenodo AfricaNLP Community**	An online catalogue that provides African language resources (data and models) in both texts and speech.	

2 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.

2.1 Data Search, Analysis, & Exploration

Exploring training datasets with search and analysis tools helps practitioners develop a nuanced intuition for what's in the data, and therefore their model. Many aspects of data are difficult to summarize or document without hands-on exploration. For instance, text data can have a distribution of lengths, topics, tones, formats, licenses, and even diction. We recommend developers use the many available tools to search and analyze their training datasets.

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. (2023).	
		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.	

2.2 Data Cleaning, Filtering & Mixing

Data cleaning and filtering is an important step in curating a dataset. Filtering and cleaning remove unwanted data from the dataset. They can improve training efficiency as well as ensuring that data has desirable properties, including: high information content, desired languages, low toxicity, and minimal personally identifiable information. We recommend that practitioners consider the possible trade-offs when using some filters. For example, Dodge et al. (2021) find that some filters disproportionately remove text written by, and about, minority individuals. Additionally, Welbl et al. (2021) and Longpre et al. (2023c) find that removing content that classifiers believe are "toxic" can have adverse affects, including lowering performance on evaluations, and disproportionately removing text representing marginalized groups. Data mixing is another important component of data preparation, where the mixture proportions of pretraining data domains (e.g. scientific articles, GitHub, and books) have been shown to dramatically affect downstream performance (Gao et al., 2020; Xie et al., 2023; Albalak et al., 2023). For more details on cleaning, filtering, and mixing (and deduplication), see a recent survey by Albalak et al. (2024).

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).	

2.3 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. Practitioners should always determine whether duplicated data will harm or help the model for their use case. For example, memorization is a crucial component for a model intended to be used in a closed-book question answering system, but will tend to be harmful for application-agnostic models (Lee et al., 2022a).

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).	

2.4 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. The following resources aid in proactively protecting test data with canaries, decontaminating data before training, and identifying or proving what data a model was trained on. Jagielski (2023) explains how to interpret canary exposure, including by relating it to membership inference attacks, and differential privacy. Oren et al. (2023) provides methods for provable guarantees of test set contamination in language models without access to pretraining data or model weights.

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.	GitHub
T	**Carper AI Decontamination Tool**	A repository, heavily based by the BigCode repository, to decontaminate evaluation sets from a text training set.	GitHub
T	**Data Portraits**	A tool to test if a model has seen certain data, e.g. in the The Pile or The Stack (Marone and Van Durme, 2023).	Paper, Web
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).	Paper, GitHub, Web

2.5 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.

At early stages of a project the data search, analysis, & exploration tools outlined in Section 2.1 are typically sufficient to track the evolution of a dataset. However it can also be helpful to do systematic studies of the process.

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

3 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 appropriately proactively organize its access, licenses, and stewardship.

3.1 Data Documentation

When releasing new data resources with a model, it is important to thoroughly document the data(Bender and Friedman, 2018; Holland et al., 2020). Documentation allows users to understand its intended uses, legal restrictions, attribution, relevant contents, privacy concerns, and other limitations. 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., 2021; Longpre et al., 2023b).

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).	

3.2 Data Governance

Releasing all datasets involved in the development of a Foundation Model, including training, 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, AmInTheStack).

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.	

4 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.

4.1 Pretraining Repositories

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., 2023b).	
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.	

Practitioners should consider using already-optimized codebases, especially in the pre-training 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.

Here, we provide a sample of existing widely-used pre-training codebases or component tools that developers can use as a jumping-off point for pre-training foundation models.

4.2 Finetuning Repositories

Fine-tuning, or other types of adaptation performed on foundation models after pretraining, are an equally important and complex step in model development. Fine-tuned models are more frequently deployed than base models.

Here, we also link to some useful and widely-used resources for adapting foundation models or otherwise fine-tuning them. Use of these tools 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.

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., 2023).	
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).	

4.3 Efficiency and Resource Allocation

Knowledge of training best practices and efficiency techniques can reduce costs to train a desired model significantly. Here, we include a select few readings and resources on effectively using a given resource budget for model training, such as several canonical papers on fitting *scaling laws*, a common tool for extrapolating findings across scales of cost. These are used frequently to determine the most efficient allocation of resources, such as allocating compute between model size and dataset size for a given budget.

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. Practitioners should thus be aware of potential second-order effects of their model releases and training choices.

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).	

4.4 Educational Resources

Training models at any scale can be quite daunting to newer practitioners. Here, we include several educational resources that may be useful in learning about the considerations required for successfully and effectively training or fine-tuning foundation models, and recommend that practitioners review these resources and use them to guide further reading about model training and usage.

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.	

5 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 3).

5.1 Estimating Environmental Impact

Current tools, including the ones mentioned in the table, focus on the latter point by measuring the energy consumed during training or inference and multiplying it by the carbon intensity of the energy source used. While other steps of the model life cycle (e.g. manufacturing hardware, heating/cooling datacenters, storing and transferring data) also come with environmental impacts, we currently lack the information necessary to meaningfully measure these impacts (Luccioni et al., 2023b). The table below outlines resources for back-of-the-envelope estimations of environmental impact, in-code estimation, as well as dashboard for cloud computing platforms to estimate environmental impact (Anthony et al., 2020; Lacoste et al., 2019).

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).	

5.2 Effective use of resources

Several decisions made during or prior to model training can have significant impacts on the upstream and downstream environmental impact of a given model. Use Scaling Laws (Kaplan et al., 2020) and other methodologies to find the best allocation of your compute budget. For models frequently used downstream, consider the inference footprint and inference cost during model creation, to minimize the environmental impact of inference (Muennighoff et al., 2023b). For further resources and discussion, see 4.3.

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.	

6 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 an evaluation 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.

6.1 Capabilities

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., 2023).	
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).	

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 which are unable to critically evaluate so many tasks, applications, and risks systematically or fairly. As a result, it is important to carefully scope the original intentions for the model, and the evaluations to those intentions. Even then, the most relevant evaluation benchmarks may not align with real use, and so should be qualified with their limitations, and carefully supplemented with real user/human evaluation settings, where feasible.

Below we note common benchmarks, as of December 2023, but caution that all of these come with substantial limitations. For instance, many multiple choice college knowledge benchmarks 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).

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., 2023).	
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.	

6.2 Risk & Harm 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 and uses 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, whereas others are focused on longer term threats related to dangerous weapons development, cybersecurity, and military use. These tend to focus on future model capabilities and their misuse (Brundage et al., 2018). Many taxonomies assess the available evidence for the risks and hazards, discuss their impact, and offer mitigation strategies (Deng et al., 2023). There is a substantial focus on text-only models and future work should consider paying more attention to multimodal models.

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.	

6.3 Risks & Harms

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., 2020).	
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 and 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).	

Evaluations of risk serve multiple purposes: to identify if there are issues which need mitigation, to track the success of any such mitigations, to document for other users of the model what risks are still present, and to help make decisions related to model access and release. Harm is highly contextual (Dingemanse and Liesenfeld, 2022; Koenecke et al., 2020), so developers should consider the context in which their foundation model might be used and evaluate the highest severity and most likely risks.

To think through the possible risks, many taxonomies of harm have been created and provide good starting points. Determining how to evaluate risk is also challenging, as there are risks and modalities with limited evaluation coverage. The sample included below are a starting point for certain key areas, but we encourage developers to browse the evaluation repository (linked below) to see if there is something more suited to their needs. In addition to fixed benchmarks, an emergent approach to evaluation is using one model to evaluate another, as done by Perez et al. (2022) and in Anthropic's Constitutional AI work (Bai et al., 2022).

7 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. For open-source licenses, there are several tools that are available to help developers select the right license for their 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.

7.1 Model Documentation

It is important to document models that are used and released. Even models and code released openly are important to document thoroughly, in order to 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.

We include frequently-used standards for model documentation as well as tools for easy following of standards and creation of documentation.

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).	

7.2 Reproducibility

Model releases often go accompanied with claims on evaluation performance, but those results are not always reproducible, or can be misleading (Kapoor et al., 2023). If code is not released, is not comprehensive, is difficult to run, or misses key details, this will cost the scientific community time and effort to replicate and verify the claims. Replication time will also slow progress, and discourage developers from adopting that resource over others.

For these reasons, we strongly recommend carefully curating code, for model training, inference and evaluation. Reproducible code begins with clear dependencies, versioning, and setup scripts, that are easy to adopt even if the tools and frameworks are unfamiliar. Clear documentation, code legibility and scripts for each entry point also improve ease of adoption. Notably, Colab Notebooks provide shareable environment setup and execution tools. These measures will significantly improves scientific reproducibility, and transparency.

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.	

7.3 License Selection

Foundation models, like software, are accompanied by licenses that determine how they may be distributed, used, and repurposed. There are a variety of licenses to choose between for open foundation model developers, presenting potential challenges for new developers. The table below includes resources that can help guide developers through the process of selecting a specific license for their model as well as several examples of licenses that include use restrictions. While licenses with use restrictions may be appropriate for certain types of models, in other cases use restrictions can limit the ability of certain categories of stakeholders to re-use or adapt the models (Foundation, 2024).

Responsible AI Licenses in particular, including BigScience's Open RAIL and AI2's ImpACT Licenses, have seen growing adoption, but also criticism of the difficulties they may pose even for well-intentioned actors seeking to comply with their requirements—especially in commercial applications—and because their enforceability still remains an open question (Downing, 2023). While they can provide a convenient way to help a developer express their understanding of their model's limitations, in conjunction with a model card that outlines in-scope and out-of-scope uses, adopters should also consider unintended consequences in limiting the scope of the follow-up research that may be conducted with the licensed artifacts. Responsible AI licenses can act as a useful norm-setting and self-reflection tool, but users should be aware of their limitations and potential downsides, especially compared to established open-source software licenses.

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.	
		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.	

7.4 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 table below includes resources related to usage monitoring, including examples of how to watermark content, provide guidance on appropriate use, report adverse events associated with model use, and 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.

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.	

Acknowledgements

Stella Biderman, Hailey Schoelkopf, and Aviya Skowron's work on this project was funded in part by a grant from the Omidyar Network.

References

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

David Ifeoluwa Adelani, 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.

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, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 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.

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

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.

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].

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

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

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*, pages 2397–2430. PMLR, 2023.

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.

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.

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, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 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.

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`.

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, page 855–868. Association for Computing Machinery, 2023. doi: 10.1145/3600211.3604658. URL `https://doi.org/10.1145/3600211.3604658`.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

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

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.

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.

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)*, pages 5614–5633, 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*, pages 1286–1305, 2021.

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, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023.

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.

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. *arXiv preprint arXiv:2009.11462*, 2020.

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.

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.

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*, pages 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.

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.

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, pages 12942–12950, 2023.

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

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)*, pages 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.

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

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

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.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

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)*, pages 215–221, 2022.

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, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.

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.

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, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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.

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*, 2023.

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

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*, 2023.

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.

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*, pages 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.

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*, pages 4969–4983, 2020.

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, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023c.

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.

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

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*, pages 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.

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.

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, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 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`.

OpenAI. Preparednessframework(beta), 2023.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*, 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`.

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.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. Rwkv: Reinventing rnns for the transformer era, 2023a.

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, 2023b.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

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.

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

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.

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.

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.

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.

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)*, pages 1–5. IEEE, 2023.

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.

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*, page 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, 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, 2022.

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.

Eric Michael Smith, Melissa Hall Melanie Kambadur, Eleonora Presani, and Adina Williams. " i'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*, 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*, 2023.

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.

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.

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.

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.

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

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.

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.

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, page 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.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, 2021.

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*, 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)*, pages 6182–6186. IEEE, 2022.

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.

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.

A Contributions

To create this cheatsheet, a variety of contributors were asked to propose resources, papers, and tools relevant to open foundation model development. Those resources were grouped into sections, which were each curated by a subset of the contributors. We list the main curators of each section, listed alphabetically below. However, it is important to note that many contributors advised across sections, and helped with preparing the interactive cheatsheet tool. Nay San led the speech modality, and Gabriel Ilharco led the vision modality.

- **Pretraining Data Sources** David Adelani, Stella Biderman, Gabriel Ilharco, Kyle Lo, Shayne Longpre, Luca Soldaini, Nay San
- **Finetuning Data Catalogs** David Adelani, Stella Biderman, Gabriel Ilharco, Shayne Longpre, Nay San
- **Data Search, Analysis, & Exploration** Stella Biderman, Gabriel Ilharco, Shayne Longpre, Nay San
- **Data Cleaning, Filtering, & Mixing** Alon Albalak, Kyle Lo, Luca Soldaini
- **Data Deduplication** Alon Albalak, Kyle Lo, Shayne Longpre, Luca Soldaini
- **Data Decontamination** Stella Biderman, Shayne Longpre
- **Data Auditing** Stella Biderman, Aviya Skowron
- **Data Documentation** Stella Biderman, Aviya Skowron
- **Data Governance** Stella Biderman, Yacine Jernite, Sayash Kapoor
- **Pretraining Repositories** Stella Biderman, Gabriel Ilharco, Nay San, Hailey Schoelkopf
- **Finetuning Repositories** Gabriel Ilharco, Nay San, Hailey Schoelkopf
- **Efficiency & Resource Allocation** Hailey Schoelkopf
- **Educational Resources** Hailey Schoelkopf
- **Estimating Environmental Impact** Peter Henderson, Sayash Kapoor, Sasha Luccioni
- **Effective use of Resources** Sayash Kapoor, Sasha Luccioni
- **General Capabilities** Rishi Bommasani, Shayne Longpre
- **Risks & Harms** Maribeth Rauh, Laura Weidinger
- **Risks & Harm Taxonomies** Bertie Vidgen
- **Model Documentation** Sayash Kapoor, Shayne Longpre
- **Reproducibility** Stella Biderman, Shayne Longpre
- **License Selection** Stella Biderman, Yacine Jernite, Kevin Klyman, Aviya Skowron, Daniel McDuff
- **Usage Monitoring** Kevin Klyman

- **Website** Shayne Longpre, Luca Soldaini
- **Advising** Stella Biderman, Peter Henderson, Yacine Jernite, Sasha Luccioni, Percy Liang, Arvind Narayanan, Victor Sanh