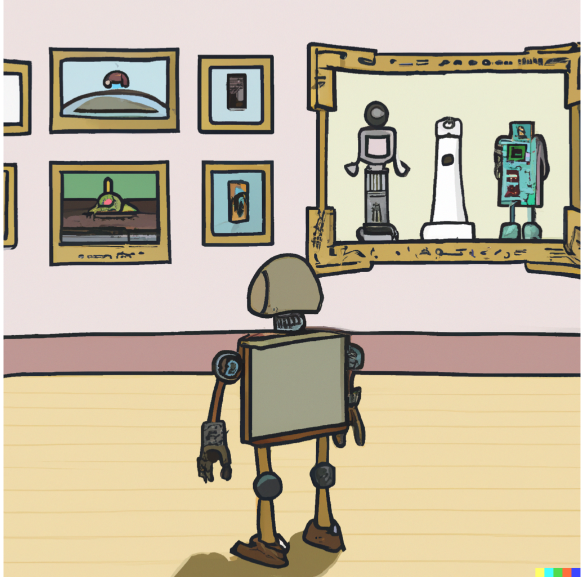
AI and Law: The Next Generation

An explainer series[1]



Katherine Lee[2], A. Feder Cooper[3], James Grimmelman[4], Daphne Ippolito[5]

July 6, 2023

[1] Available in html at: https://genlaw.github.io/explainers/
[2] https://katelee168.github.io/
[3] https://afedercooper.info/
[4] https://james.grimmelmann.net/
[5] https://daphnei.com/

Contents

Chapter 1

Introduction

by A. Feder Cooper and Katherine Lee

It's impossible to read about technological innovation right now without hearing the term "generative AI." We are in a moment of seemingly nonstop excitement (and seemingly nonstop lawsuits) about the future of AI-assisted content creation, and the questions such creation raises about data ownership, privacy, the future of work, and how technology shapes individual and collective rights. And, of course, when considering questions about rights, it is important to think not only about novel technical developments, but also novel issues such development presents for the law.

Challenging questions at the intersection of technology and law are not new. Nevertheless, recent generative AI capabilities have been so unexpected and transformative that many have been questioning if (and how) the law may need to transform in order to contend with generative AI's broader societal impact. We are concurrently seeing two highly specialized fields of knowledge undergo immense changes, with numerous opportunities for both to inform each other.

One area of particular interest is the relationship between generative AI and copyright law, especially in the context of large language models (LLMs) and diffusion-based image generation models (Clarkson, et al. v. OpenAI, et al.; GitHub Copilot litigation; Paul Tremblay and Mona Awad v. OpenAI, et al.; Stable Diffusion litigation). Systems like ChatGPT and Stable Diffusion exhibit impressive capabilities; however, they have also been shown to regurgitate training data examples in their outputs, bringing about concerns regarding infringement of intellectual property rights.

In such an ever-shifting landscape, the only certainty is that the future is uncertain. Even so, it's clear that developing expertise in either area requires being attentive to the other. Doing work in generative AI without at least a passing familiarity in copyright is increasingly intractable — and vice versa.

While comprehensive expertise in both areas is an elusive goal, it's still important to be familiar with concepts in generative AI and copyright. Familiarity with key ideas across both disciplines is essential for asking more precise questions at their intersection — questions that can meaningfully shape the futures of technical research, and law and policy.

What our explainer series aims to do

Before we can discuss precise questions at the intersection of generative AI and copyright law, we first need to develop a common understanding of some of the building blocks in each discipline. Our explainer series will provide salient details from both areas at (what we hope is) the right level of abstraction. After reading this series, ML researchers and practitioners should have a better understanding of how copyright concerns may impact their technical work, and legal experts should have a better understanding of how specific technical aspects of generative AI are important to consider when analyzing concrete implications for copyright.

What our explainer series doesn't do

This explainer series is not a machine learning paper. We don't present novel technical results on generative AI or new model evaluation metrics, nor do we aim to write a comprehensive lit review of generative AI (the pace of the field makes that impossible). We describe core concepts, such as training data, copyright, and prompting. While details may change over time, we focus on concepts that are likely to remain primary players for the foreseeable future.

This explainer series is not a law review paper. We don't provide an in-depth analysis of the implications of generative AI for copyright law. We present the contours of important concepts in copyright law, give an intuition for why they're relevant to current discussions of generative AI, and suggest connections between these concepts and important questions about evaluation of generative AI systems. Rather than doing a deep dive on a specific copyright concept (e.g., fair use), we hope that our series will give others the necessary background to be able to explore specific concepts with greater precision.

We've divided the explainer series into 4 parts

1. ***Training data***: We describe what training data is and how it is collected, putting collection processes for generative AI in historical context with prior image and text generation systems. Training datasets are created objects; we emphasize the associated choices that data collectors make, which impact trained model behavior.

2. ***Copyright***: There are currently a lot of concerns about the interplay between model behavior and copyright law. For example, there is an active debate over whether training on copyrighted data constitutes infringement or whether producing an output generation that looks almost identical to a training data example constitutes infringement. To understand these potential issues better, it's necessary to have some background information on what copyright law is and what ownership rights it is intended to protect (and what it doesn't). We provide a brief sketch of key concepts helpful for understanding why copyright is such a prevalent concern in news and lawsuits regarding generative AI.

3. ***Training models and generation***: While we primarily situate our discussion of copyright in relation to training data, other aspects of a generative AI system may implicate legal issues. We describe key terms and concepts in the process of training models and generating outputs, which rely on our prior discussions of training data and copyright.

4. ***Looking ahead***: The three posts above provide high-level background on generative AI and copyright concepts that (if we've done our jobs) should bring to light more precise discussion about emerging issues at their intersection. We describe some current trends in research at this intersection, and possible future directions.

Dedication

Chapter 1 is dedicated to the late Chris Cieri, director of the Linguistic Data Consortium, with whom we had discussed the early versions of this chapter in 2021.

Acknowledgements

This explainer series was fueled by years of discussions with wonderful people, including, but not limited to: James Bradbury, Nicholas Carlini, Chris Cieri, Lillian Lee, Shayne Longpre, David Mimno, Ludwig Schubert, Florian Tramèr, and the Artificial Intelligence, Policy, and Practice initiative at Cornell University.

Cover Image

DALL-E generation prompted with "cartoon robot looking at art in a museum."

Chapter 2

The Devil is in the Training Data

by Katherine Lee, Daphne Ippolito, and A. Feder Cooper

2.1 Introduction

The process of training contemporary generative models requires vast quantities of *training data*. Dataset creators and curators make extensive decisions about how much and which data to include in a training dataset. These choices directly and significantly shape a model's outputs (a.k.a. *generations*), including the model's capacity to learn concepts and produce novel content.

Given the sheer amount of training data required to produce high-quality generative models, it's impossible for a creator to thoroughly understand the nuances of every example in a training dataset. It's impossible for them to interact with each item in the dataset, nor can they know exactly the content, source, and context of each item in the dataset.[1] As a result, not only do their curatorial choices affect generation quality, they can also have unintended consequences that implicate legal concerns.[2] For example, generative models have been shown to generate text from copyrighted books (Wallace et al., 2020), reveal API keys (Mohammed, 2021), contact information (Carlini et al., 2021), and images with trademarks (Deutscher, 2023).

The inability to exhaustively inspect every training-data example is not specific to generative AI, nor is it a new problem. From the advent of the "Big Data" trend of the last few decades, comprehensively understanding datasets has proven to be a difficult and elusive challenge. The ways that researchers have approached this challenge are instructive for understanding contemporary practices in dataset creation and curation for generative AI.[3]

In this chapter, we begin with a history of datasets used in generative AI to understand how incentives, compute, and model design have impacted dataset development, then discuss how datasets and dataset collection practices have changed over time. We loosely trace a common pattern in both text and image models: early work on manually-constructed systems that did not ingest any training data; a transition to learning models from hand-annotated data compiled from public domain sources; and, the modern tactic of scraping massive amounts of unlabeled data from across the web. In light of

[1] For this reason, Bender et al. (2021) argues that datasets should only be as large as is possible to document.

[2] We should couch this by saying that training dataset design is the *current* most important set of choices or which model creators have to deal with training-data-based legal concerns. Other models under development are trying to reduce these risks by attributing generations to specific examples in the data, by adding noise to obscure individual data points (i.e., differential privacy), or by limiting the scope of a model to an application where copyright and privacy are less of a concern (e.g., Disney training a model on screenplays for which its own all of the relevant copyrights).

[3] Researchers' assumptions and norms, while perhaps fixed as a cultural practice (Bowker and Star, 2000; Couldry and Hepp, 2017), are not technical requirements. There are other ways that model creators could collect and curate (meta)data that would have a marked change on these assumptions and their consequences. We will return to these possibilities later.

Figure 2.1: The Utah Teapot was one of the first digital 3D models of a real-world object. Early work in computer graphics sought to render 3D objects like the teapot realistically in 2D images. (Source: "CreativeTools.se - PackshotCreator - 3D printed colourful Utah teapots" by Creative Tools is licensed under CC BY 2.0.)

this most recent approach, we discuss the choices dataset creators make when building modern-day, generative-AI datasets. Finally, we acknowledge both the difficulty in making educated choices and the impact those choices have on the resulting models.

2.2 A Brief History

While modern machine learning-based generative AI uses statistical methods to learn patterns from data, for most of the history of Generative AI, researchers did *not* use datasets in this way. Rather, early researchers built algorithms, which manually encoded patterns that allowed images and text to be generated according to the pattern. For example, early chatbots, such as ELIZA (Weizenbaum, 1966) and ALICE (1995), and early developments in novel (Klein et al., 1973) and story (Turner, 1993) generation used techniques from classical artificial intelligence to generate text based on hand-crafted rules and grammars.[4] Similarly, early work in the field of computer graphics on photo-realistic image generation focused on constructing mathematical models of 3D objects, such as the famous Utah teapot (Computer Graphics - The University of Utah, 1975), and then rendered them as 2D images. This work developed algorithms to mimic the shading and light effects of the real world, some of which were grounded directly in mathematical models from physics and optics. Other work used procedural algorithms (Wikipedia, 2023) to generate realistic textures and add them to surfaces.

2.2.1 Language Datasets

The earliest learned models for language generation built off of datasets developed by academic researchers for natural language processing (NLP) tasks — e.g., early monolingual datasets like the Brown Corpus (Francis and Kucera, 1979) and the Penn Tree Bank (Marcus et al., 1999). These early research datasets tended to be collected from literary, government, and news sources, and were densely annotated with linguistic structure, such as parts-of-speech[5] and syntax[6] annotations.

Early work in NLP assumed that building a language understanding system would require encoding this type of linguistic knowledge and mechanically applying it.[7] Building annotated datasets was a labor-intensive process and was often completed by professional, highly skilled annotators at organizations like the Linguistic Data Consortium.[8] Then, as the internet grew and expanded, training datasets began to leverage the corresponding growth in the number of electronic and

[4] The rule-based machine translation systems that pre-dated statistical machine translation (Google Operating System, 2007) are good examples of this approach..

[5] Whether a word is a noun, a verb, an adverb, etc.

[6] The hierarchical structure of words in a sentence.

[7] For example, a model might use linguistic structure to understand that in the sentence "The dog fetched the ball.", "the dog" is a noun phrase serving as the subject of the sentence. Additionally, the model might infer from context that a dog is the sort of thing that is more likely to fetch than a ball, and that a ball is the sort of thing that a dog might fetch.

[8] Some datasets, such as the Penn Treebank, were based on crowdsourced annotations by non-experts.

digitized records. Some notable datasets include English Gigaword (Graff and Cieri, 2003), sourced from news articles in English; the Enron Emails dataset (Cohen, 2015), sourced from emails released by the US federal government during its investigations of Enron's massive accounting fraud; and the One Billion Word Benchmark (Chelba et al., 2013), sourced from government documents and news.

Most of the datasets we've mentioned so far consist of material that was either a matter of public record or explicitly licensed for research use. However, rapid technological developments, which both demanded and facilitated the use of increasingly larger amounts of data, put pressure on expanding to other data sources. For one thing, it became apparent that bigger datasets led to superior language models. For another, novel algorithms and advancements in computing hardware made it possible to process datasets at previously unprecedented scales and speeds.

Efforts to build and maintain responsibly-sourced and hand-curated datasets could not keep apace with these changes. Neither could the production of manual data annotations; however, as discussed below, this presented fewer problems than anticipated, as larger models trained on larger datasets proved able to automatically pick out patterns without such curated information. Machine translation provides one of the earliest examples of generative applications to work with such large text corpora. Dataset creators assembled datasets of texts in two or more languages on the same topics from multilingual data sources, such as United Nations documents and news sites. Some of these datasets had aligned text (i.e., a specific sentence in multiple languages), but many others, like Europarl (Koehn, 2005), were simply transcripts of the same parliamentary meeting in multiple languages. These datasets were used to build statistical language models which would learn to output translations for any input sentence.

The 2010s saw a further shift from *public domain* data to web-scraped datasets compiling *publicly available* data, which can exhibit varying types of copyrights.[9] Some examples of these web-scraped datasets include the Book Corpus (Zhu et al., 2015), which scraped 11k books from Smashwords, a website for self-published e-books„ and the WritingPrompts dataset (Fan et al., 2018), which scraped the r/WritingPrompts subreddit.[10] Other datasets included data scraped from crowd-sourced platforms, such as Wikipedia and OpenSubtitles, for which it can't always be validated that the user has ownership rights over the content they upload.

For the most part, these datasets were not annotated with the rich linguistic information that accompanied older datasets. Not only were these annotations infeasible to collect on massive datasets, but advancements in machine-learning methods made them unnecessary for strong performance on many language tasks.[11]

These 2010s-era typically datasets collected documents from a single website. In contrast, more recently we have seen the growth of datasets that instead attempt to sample from the entirety of the web. Some prominent examples are RealNews (Zellers et al., 2019), C4 (Raffel et al., 2020),[12] and WebText (Radford et al., 2019). All current state-of-the-art large language models are trained on datasets scraped broadly from across the web. Separately, many companies maintain databases of data their users generate. Some of these companies have released samples or subsets of these datasets for external use. Such datasets may be annotated with user actions. For example, Amazon's Review dataset, released in 2018, contains 233.1M examples with customer ratings (Ni et al., 2019),[13] and Netflix's recommendations dataset contains 100M customer ratings (Netflix, 2009). Other popular

[9] *Public domain* typically refers to government records (such as Europarl) or works for which copyright protections have lapsed or expired. Judicial opinions, for example, are in the public domain. They are not copyrightable, which means that anyone can copy them for any purpose, and renders moot questions of copyright pertaining to content generated from such records (Wheaton v. Peters). *Publicly available* data, in contrast, is widely available but may have legal restrictions that purport to limit certain rights to certain users. For example, fanfiction uploaded to An Archive of Our Own can be freely read by anyone, but its authors-slash-users retain the copyrights in their works.

[10] Reddit recently changed its terms of service, partially in response to large generative AI companies scraping its website for training data (Ajao, 2023; Witkowski, 2023). Twitter also recently rate-limited its platform, ostensibly in response to web-scraping (Reuters Staff, 2023).

[11] Up until 2016, Google Translate used phrase-based translation, which broke sentences into linguistic parts to translate separately. These techniques have since been replaced by neural networks (Le and Schuster, 2016).

[12] See also the C4 online explorer.

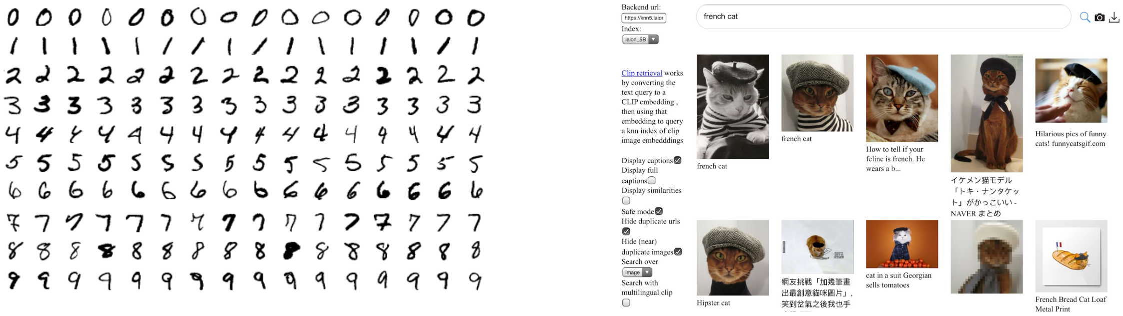
[13] See also the accompanying website.

datasets in this vein are IMDb movie reviews (Maas et al., 2011) and the Google Books N-gram corpus (2.2 TB of text!) (Google, 2012).[14]

2.2.2 Image Datasets

Image datasets have followed a similar overarching trajectory — from not using training data, to academic datasets constructed from public domain information, to industrial labs building massive datasets scraped from the web. Until relatively recently, most image datasets were developed with the goal of producing applications that annotated or classified images, rather than generating them. Early datasets include MNIST, which consists of 60,000 black-and-white images of handwritten digits (LeCun and Cortes, 1999); CIFAR-10, which contains 60,000 photographs of objects from 10 classes, including airplanes, frogs, and cats (Krizhevsky, 2009); and ImageNet, which has over 14 million images divided among 1,000 classes (Deng et al., 2009). Deep-learning researchers relied heavily on these datasets to develop methodologies for image classification, and early work on machine-learning-powered image generation, including generative adversarial networks (GANs) (Goodfellow et al., 2014) and denoising diffusion models (Ho et al., 2020), followed suit.

For many years, image generation models were not powerful enough to capture the diversity of all image classes represented in a dataset like ImageNet.[15] Thus, more narrowly-scoped (but still large) datasets allowed for the development of models with higher-fidelity generations. Prominent examples include Celeb-A with 200k close-up images of faces (Liu et al., 2015); Caltech-UCSD Birds with over 11k bird photos (Welinder et al., 2010); and, Oxford Flowers 102 with 1k flower photos (Nilsback and Zisserman, 2008).



Backend url: https://knn5.laior

Index: laion_5B

french cat

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑

Display full captions ☐

Display similarities ☐

Safe mode ☑

Hide duplicate urls ☑

Hide (near) duplicate images ☑

Search over image

Search with multilingual clip ☐

french cat

french cat

How to tell if your feline is french. He wears a b...

イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ

Hilarious pics of funny cats! funnycatsgif.com

Hipster cat

網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手

cat in a suit Georgian sells tomatoes

French Bread Cat Loaf Metal Print

Figure 2.2: **left**: Examples from the MNIST dataset (LeCun and Cortes, 1999). **right**: Examples from LAION-5B (Schuhmann et al., 2022) (Screenshot by the authors). Note the uniformity of the images in MNIST vs. the varying aspect ratio and image quality in LAION. Additionally, LAION images come with much longer captions that don't always exactly describe what is in the image, whereas labels associated with MNIST images are the number written in the image.

Of course, the most exciting models today don't just generate an image from a single class identifier like "bird" or airplane." Models like Stable Diffusion or Imagen can parse complex natural-language descriptions to generate novel and complex images in a variety of styles.

Good text-to-image models require the use of training images with rich captions describing them. Early GAN-based text-to-image models used datasets with human-written descriptions of the images in the Birds and Flowers datasets (Reed et al., 2016). For this reason, datasets intended for image captioning research, like MS-COCO (Lin et al., 2014), were also used in reverse for this purpose:

[14] Some of these datasets contain data from sources with mixed ownership, and thus have been subjected to copyright claims. Notably, authors and publishers sued Google for copyright infringement over its collection and use of scanned books. Google ultimately prevailed, following a decade of litigation (Authors Guild v. Google).

[15] This is not in contradiction with the earlier point that larger training datasets tend to produce better models. Rather, it is the case that larger, more capable models enabled more effective utilization of large datasets.

Such datasets had their examples flipped — i.e., using the caption labels as training data, and the image as label — so that the resulting models could generate an image from a caption.

Just as with language research, it quickly became clear that larger datasets resulted in superior models.[16] LAION-5B (Schuhmann et al., 2022) met this need with a dataset of over 5 billion image-text pairs, created by extracting images with detailed alt-text descriptions (Accessible Publishing)[17] from the [Common Crawl](https://commoncrawl.org) web-scraped corpus. LAION-5B is the dataset on which most state-of-the-art open-source text-to-image models are trained.

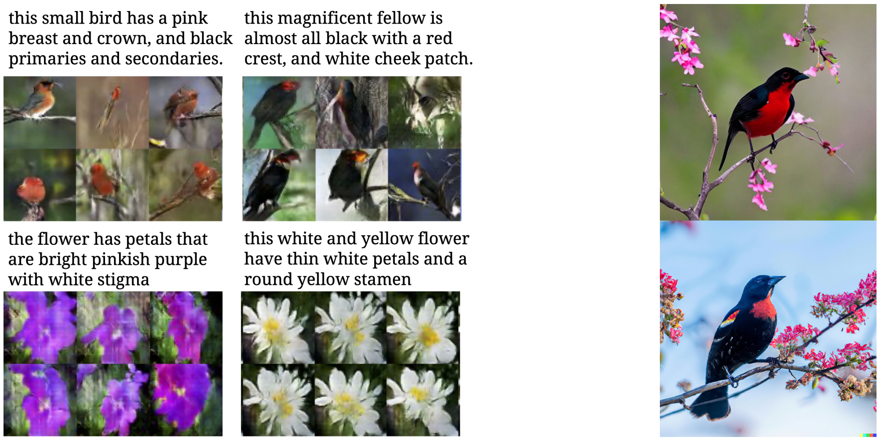

this small bird has a pink breast and crown, and black primaries and secondaries.
this magnificent fellow is almost all black with a red crest, and white cheek patch.
the flower has petals that are bright pinkish purple with white stigma
this white and yellow flower have thin white petals and a round yellow stamen

Figure 2.3: **left**: Generations from one of the first text-to-image synthesis papers (Reed et al., 2016, Figure 1). **right**: Images from the state-of-the-art text-to-image models DALL-E 2 (OpenAI, 2023) and Stable Diffusion (Stability AI, 2022)) for the prompt "A red-breasted black bird perched on a branch with small pink flowers" (prompted by the authors). Modern models are much better at composition and synthesizing novel scenes than older systems.

2.3 Today's Datasets

As discussed above, the datasets used to train today's large language models are massive and predominantly contain data scraped from the web.[18] Among popular language datasets, The Pile (Gao et al., 2020) and C4[19] (Raffel et al., 2020) are both 800GB; ROOTS is 1.6TB of pre-processed text (Laurençon et al., 2023), and Chinchilla (which is not publicly released) is 5.3 TB.[20] These datasets are of a completely different scale than those mentioned in prior sections, and in turn present unprecedented challenges for data maintenance and curation. In this section, we describe some of these challenges — the ramifications of datasets that are orders of magnitude too large for their creators to manually inspect each data point. We'll compare modern datasets with the MNIST image dataset (mentioned above), the quintessential small dataset in machine learning.[21] In doing so, we emphasize that datasets are manufactured objects; there are numerous choices that dataset creators and curators make in the production and maintenance of datasets.

[16] In addition, training datasets needed to be high-resolution in order to create high-resolution, high-fidelity generations.

[17] Alt-text descriptions of images are an accessibility feature intended for situations where an image cannot be rendered. For example, visually impaired people using screen readers will read alt-text in lieu of seeing the image.

[18] The Pile, ROOTS, and Chinchilla all combine data scraped from the web with additional more-curated data sources.

[19] See also the Hugging Face data card.

[20] The number reported in their paper was 1.4T tokens (Hoffmann et al., 2022), or 4x the training data for a different language model, Gopher, which used 300B tokens (Rae et al., 2022) FThe Gopher creators sampled 12.8% of the MassiveText dataset, which contains 10.5TB of data. 0.128 * 4 * 10.5TB = 5.3TB.

[21] Over 6000 papers cite MNIST directly and a Google Scholar search returned 76,900 articles that mention MNIST. MNIST has become a generic term for "small, standard dataset" so other datasets like Fashion-MNIST, which has photos of clothing, also appear in the search results.

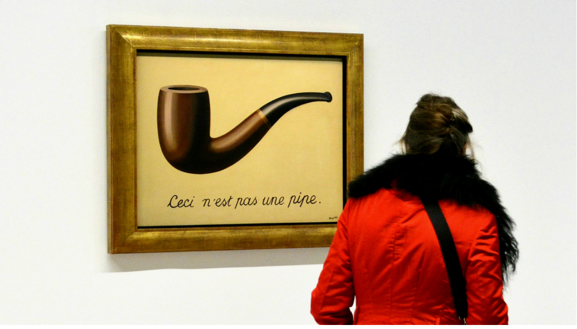


Figure 2.4: Est-ce French? Is this l'anglais? (Source: "Looking at Magritte I" by C. B. Campbell is licensed under CC BY 2.0.

2.3.1 Choosing Training Data Sources

Dataset creators choose data in two stages. The first consists of picking whether or not to include entire sources of data, and the second includes applying automatic filters to remove individual unwanted examples. Both these steps can involve creators making dozens of choices about what data is relevant.

Data scraped from one source is typically called a *corpus*. Corpora may be scraped from Twitter, code repositories like Github, personal blogs, advertisements, FanFiction, PasteBin dumps, search-engine optimization text, and so on; image datasets can come from data aggregators like Flicker, Shutterstock, Getty, or be gathered from a crawl of the entire web.

In choosing one corpus over another, dataset creators make assumptions about the content of each one. For example, if the dataset creator wanted to create a model that was able to give coding advice, they might choose to include Stack Exchange or GitHub data as well.[22] Alternatively, Wikipedia is generally a popular source of data because it contains curated articles on a diverse array of topics.[23] For both image and language datasets, creators can also make the decision to scrape from a crawl of the entire Internet, rather than specifically seeking out certain websites or domains.

The composition of languages within a dataset is another important question. Should the dataset be primarily one language? Should it include as many languages as possible? If so, should the distribution of examples be balanced equally? Each answered question leads to even more choices.

If an English sentence includes a single Italian word, is that sentence English or Italian? What about the sentence, "I walked from campo dei fiori to santa Maria degli angeli?" Additionally, many uses of language are contextual and cultural. Before René Magritte's 1929 painting *The Treachery of Images*, one would have said "Ceci n'est pas une pipe" was French, but today, it would also be commonly understood by many English speakers. Further, different languages vary with respect to how they convey similar ideas; German uses of long compound words to reflect complex concepts, while other languages like Japanese may only use a single logogram.

We know that the balance of genres in a dataset affects the resulting model's knowledge and abilities. A model trained on Project Gutenberg, a collection of public-domain eBooks whose most recent entries were first published in the 1920s, will clearly be much worse at reciting recent facts than one trained on Wikipedia. While this example might seem intuitive and obvious, we mostly don't understand with specificity how upstream dataset collection choices affect downstream generations. As much as we would like to make each data-selection decision scientifically, on the basis of good evidence and a careful weighing of competing goals, it is cost- and compute- prohibitive to run a

[22] GitHub contains far more than just code. For example, many repositories also contain READMEs written in prose. Additionally, since many websites and blogs are hosted on GitHub, GitHub can also contain personal, narrative stories.

[23] Because of this diversity, Wikipedia may be included when training wide-purpose applications like chatbots. However, Wikipedia isn't conversational. So, for interactions with the generative model to feel fluid and natural, the dataset creator may choose to also include chat data, such as YouTube subtitles, HackerNews conversations, or Twitter free-for-alls.

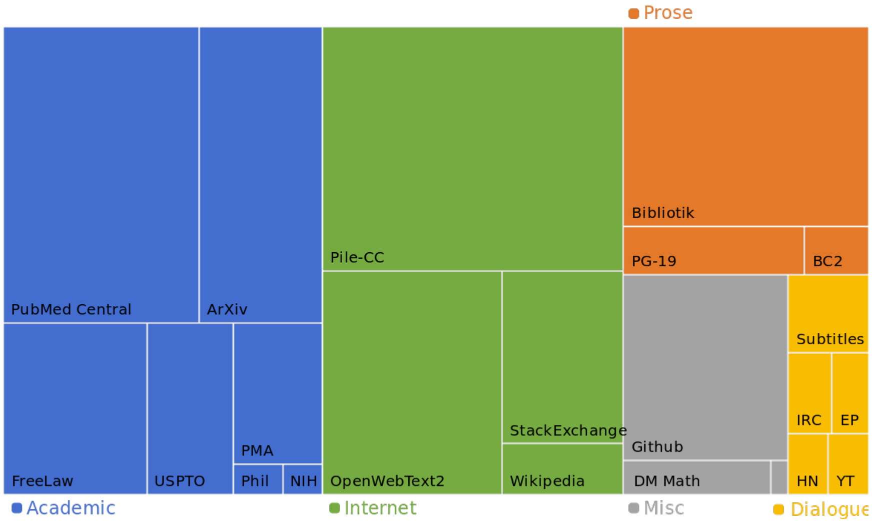


Figure 2.5: The Pile (Gao et al., 2020) is made up of many smaller datasets. Many of these components are web-scrapes focused on a specific domain, such as Wikipedia, StackExchange, USPTO (United States Patent and Trademark Office), and arXiv. Some components, like Enron Emails (Cohen, 2015), EuroParl (Koehn, 2005), and Project Gutenberg (a collection of out-of-copyright books available online) are not explicitly scraped from the web. Figure produced by Ludwig Schubert (adapting Figure 1 from Gao et al. (2020) to create a .svg version).

different experiment for each decision. Thus, many of these decisions are simply choices the dataset creator makes, without much validation.

As a concrete example, we can study The Pile (Gao et al., 2020), a popular publicly-available dataset for training language models. The Pile's creators chose to include multiple "academic" datasets, like PubMed, arXiv, and FreeLaw, and code from GitHub. This means that models trained on The Pile will have seen medical literature, legal literature, and code. A model not trained on code would have a much harder time generating it.[24]

2.3.2 Identifying "Good" Data Examples

While dataset creators frequently say they want "clean data," the term is a misnomer. Instead, dataset creators typically mean that they want a dataset that creates a "good model." Defining what "good" means similarly involves dataset creators to make many choices.

This is true even for models with seemingly clear goals, for which "good" can seem easy to define. The creators of MNIST wanted a model that could classify images of handwritten digits, and the dataset they collected could be tailored to this single, specific goal. However, models trained on MNIST can (and have been) used in a variety of different tasks, with varying degrees of success. For example, MNIST-based models have been useful for digitizing zip codes on postal envelopes; however, they may be less useful when applied to writing that is handwritten or drawn in more stylized writing, e.g., in artwork.

Defining "goodness" is even more difficult for generative AI, in part because generative AI is significantly more flexible (literally, generative) than traditional classification tasks that have clear output labels and single, unambiguously correct answers. This flexibility is a desirable feature for generative AI; we want our models to do many things. For example, we expect a large language

[24] This doesn't mean that models not explicitly trained on code can't generate code. Webscraped data is not cleanly separated into different semantic domains, and there will inevitably be some code mixed in with whatever text the model is trained on.

model to be able to answer factual questions about the world, but also to tell fictional stories, give relationship advice, and de-escalate when asked to generate toxic content.[25]

The multiplicity of uses (many of which are yet to be determined) means “good” is extremely underspecified, and thus many different choices of training datasets could be “good” in different ways.[26]

2.3.3 Filtering Out “Bad” Data Examples

“Good” data is hard to define, and “bad” data is similarly ambiguous. To be a little more specific, consider “toxic content” as an example of “bad” data for training a conversational LLM meant for wide public use. We might want to filter out “toxic content” from the data we scrape and not include it in our training dataset — an easy goal to state but hard to implement.[27]

“Toxic content” is ill-defined[28] and constantly evolving. Metrics of toxicity can be correlated with other aspects of text, such as sexual explicitness.[29] For example, the Texas Liberty County Vindicator posted the full text of the Declaration of Independence and Facebook’s moderation flagged it as hate speech (Rosenberg, 2018). Additionally, different individuals or groups may have different interpretations of the same text, complicating the process of deciding what data to exclude.[30]

Further, even if we agree on what is “toxic content,” there unfortunately isn’t a clear consensus on whether excluding such content from the training data is an effective strategy for preventing it from being generated. While some researchers propose removing any data deemed “toxic,” others disagree. They believe a better strategy is to control model outputs, not inputs; in their view, including some “toxic” data helps models to identify it and thus stop its generation (Longpre et al., 2023).

Identifying “toxic content” (and including or excluding it) is only one of many classes of many, many choices that dataset curators have to *just make*. And no set of choices is every complete. Any fixed, black-and-white process of determining what data is worth including or excluding will miss cultural connotations that resist quantification and objectivity. Whatever processes we use for deciding on “good” and “bad” data must be adaptable and open to revision as society and model uses evolve.

Moreover, the scale of today’s datasets encourages – indeed, requires – dataset creators to use automatic methods to decide what data to include or remove. For “toxic content,” it is common practice to use a classification model (typically one that is small and fast to run) to determine what to exclude. This classification model might have been built using human annotations, or it might itself be built with automatically derived labels.[31] The model is then used to automatically

[25] At least, this is true right now. It is possible that, over time, it may be desirable for a chatbot to exhibit more narrow functionality: A chatbot that supplies financial advice may be subject to regulation, and it may not be appropriate for it to also give relationship advice.

[26] Many generative AI models are referred to as “general purpose” (this is the G and P in OpenAI’s GPT). Models are in fact never completely general because data and modeling choices create preferences and limitations. However, the intent of some creators is to make the model as general as possible. When model creators say “good” they sometimes mean they want “helpful and harmless models” (Bai et al., 2022). Sam Altman also used this phrasing during the Senate Judiciary Committee hearing on AI. The complex relationship between “general purpose” models and specific end uses also predates the recent uptick in generative AI (Cooper et al., 2022).

[27] Dataset creation and collection choices are entangled with overall learning goals. Even if we have a fixed and agreed-upon definition of what content is “toxic,” a model trained on a “good” dataset that contains no toxic content may be “good” in that that it does not generate toxic content on its own, but fail to be “good” in that it does a poor job of summarizing and explaining a user-provided article that contains toxic content. The point is that “good” is a slippery and can mean different things in different parts of the generative AI pipeline. This is also why “helpful” and “harmless” are sometimes linked as joint goals; they are both important, but they are distinct and can be in tension with each other.

[28] Antoniak and Mimno (2021) demonstrate how measurements of bias can themselves be biased based on choices of what topics to measure.

[29] The Perspective API tries to identify “toxic” content (Jigsaw, 2017), best understood here as “stuff you don’t want advertisements associated with.”

[30] For example, Dodge et al. (2021) discusses how the collection process for the C4 dataset disproportionately filters out data related to certain demographic groups. One way to approach this challenge is to adopt a more flexible and inclusive approach to filtering criteria and analysis. Overall, it is important to recognize that cultural data is often fluid and dynamic, and our understanding of it may change over time. Therefore, any process for determining what data to include and exclude must be adaptable and open to revision as new insights emerge.

[31] A popular approach right now is to train a classifier that assesses quality using a training dataset where examples

label every example in the dataset, and examples with negative labels are removed. For example, LAION-Aesthetics (Schuhman, 2022) is a subset of LAION-5B containing only images that an automatic classifier labeled as "aesthetically pleasing." Such an automated process also contains assumptions and choices, and likely does not perfectly capture nebulous concepts like "toxicity." It's training data choices all the way down.

Other difficult and contestable choices around dataset curation have to do with the possibility of errors. Datasets where each item is labeled can contain errors because there is more than one way for an example to be labeled, resulting in misleading or incomplete labels.[32] In image-caption datasets, an image's caption may describe only one part of the image, or it may not even correctly describe the content of the image at all. In any language dataset, the text might not be in the language we expect it to be, or it could be written in multiple languages, or it might not even be natural language.[33] The massive size of today's datasets makes it extremely challenging to systematically identify and remove examples with these types of errors.[34]

That being said, modern models are remarkably capable of performing tasks *in spite* of misleading or mislabeled examples (Zhang et al., 2021). Generative language models are typically trained with the objective of predicting the next word in the sentence given the previous ones[35] and are used to perform tasks they weren't explicitly trained to do. None of the current language models were explicitly trained to answer questions about Dolly Parton, but they will deliver topical and appropriate responses to those questions. Additionally, modern models can also perform mechanical tasks like reversing a sentence.[36]

2.3.4 Testing Training Data Choices

Ideally we would like to know exactly dataset creation choices will impact the generative model. The standard approach to testing this is to train models on different slices of training data, then evaluate how the removal of each slice impacts the resulting model's performance. This approach is called *ablation testing*. As a concrete example, we could train the same model with and without Wikipedia in the dataset, or with and without text classified as "toxic," and then observe how well the resulting models perform on tasks like question-answering and "toxic" tweet identification.

Unfortunately, ablation testing is prohibitively expensive in contemporary generative AI. Today's models are massive (billions of parameters) and can cost millions of dollars to train, and therefore are typically only trained once (or a small handful of times) on the whole training dataset. While it could be feasible to do ablation testing with smaller models, they don't always yield the same results as testing on larger ones. Model creators simply can't afford to test every possible definition of "toxic" or every combination of "include/exclude" for different types of data.

It is worth emphasizing that training models at such large scales is also a choice. We could choose to train smaller models on smaller datasets for which ablation tests are tractable. But, doing so would sacrifice the powerful generalization capabilities of large-scale generative AI.[37] An important takeaway, though, is that running one pass on a giant training dataset (i.e., *not* being able to do ablation testing or other procedures that involve multiple training runs, like hyperparameter optimization) is

are labeled as high-quality if they come from a trusted source like Wikipedia or Books, and low-quality if they come from the general Internet.

[32] For example, an analysis of ImageNet's biodiversity found that 12% of the dataset's wildlife images are incorrectly labeled (Luccioni and Rolnick, 2022).

[33] One of the authors of this paper spent days confounded by why a model trained on Project Gutenberg (Project Gutenberg) was generating gibberish. It turned out the gibberish was chess moves, and that 18 million characters of chess notation were in the dataset (Fishburne, 2003).

[34] Just consider that MNIST, with 60,000 examples is 97,500x smaller than LAION, with 5.85 billion.

[35] Some modern LMs are trained with other objective functions, such as a fill-in-the-blank-style objective called span corruption (Tay and Dehghani, 2022). Many generative image models are trained to reconstruct a given training example.

[36] Try it yourself in ChatGPT. Providing examples of what "reversal" means in the prompt to the model can help the model understand the pattern, but models are not successful every time and are very sensitive to the format of the prompt.

[37] For a while, it was understood that models develop capabilities at larger sizes (Wei et al., 2022), though recently Schaeffer et al. (2023) challenged this understanding.

not a foregone conclusion. This reflects choices in what developers and researchers have opted to prioritize for developing generative AI.

2.3.5 Understanding Provenance

Automated data collection processes can obscure provenance. For large, web-scraped datasets, dataset creators might know that an image or a paragraph of text is from a particular website. Unfortunately, website origins don't necessarily correlate with authorship, so that content's presence on a website doesn't prove that the website had permission to post the content. For example, chat logs are typically between two or more people. Both people don't need to consent for one person to put that chat log somewhere public, where it could become part of the dataset. The chats could also become public through a data leak, or as a result of malicious action. For example, Sony executives' emails were stolen and posted by North Korean hackers, and personal chats were leaked during the GamerGate harassment campaign.[38] The entire movie of *The Fast and the Furious* could become part of a dataset without the dataset creator's knowledge because a Twitter user decided to tweet out the entire movie in two minute clips (Townsend, 2022).[39]

To be clear, this reality is also a reflection of choices in contemporary dataset practices. Older datasets tended to be more curated and constructed, which made the provenance of the data in them clearer. For example, the MNIST dataset was built by *M*ixing two datasets from the *N*ational *I*nstitute of *S*cience and *T*echnology: one of handwritten numbers, one written by high school students, and the other by employees at the US Census Bureau. The provenance of each digit was clear.

Instead, with the choice to move away from manual curation toward scraping massive datasets, provenance has become harder to track and understand.[40] This can cause a host of problems, such as issues around attribution. Additionally, without provenance and inferred cultural context, data may look unexpected. For example, a generative AI model may refer to "sex" as "seggs" because individuals online have adapted to censorship by using homophones like"seggs" to discuss sensitive topics.

2.3.6 Using Data that We Don't Fully Understand

As datasets are used more, our understanding of them improves. Older datasets have been around long enough for researchers to develop an understanding of their flaws. For MNIST, we even know how many and which examples are labeled incorrectly.[41]

The rapid pace of generative AI research makes it difficult for analysis of existing datasets to keep up with the development and adoption of new ones. This is also a choice; developers could choose to wait to study a dataset more carefully before training with it.[42] Additionally, an increasing number of popular generative models are trained by companies on non-public datasets — making outside analysis impossible. For example, we don't know much about the training data for ChatGPT, nor the difference between ChatGPT's and Claude's training datasets. However, to the extent that similarity between training data and the user's downstream task has an impact on the generative AI's performance on the tasks, companies should feel motivated to document and release additional information about what was in the training data to enable users to choose the right API for their application.

[38] GPT2, another language model, generated a conversation between two real individuals using their usernames. This conversation wasn't exactly as it appeared in the GamerGate harassment campaign, but was about the same topic. More on this subject in Wallace et al. (2020)'s blog post and in Brown et al. (2022).

[39] Another example of a copyright concern involves license laundering (Wikipedia, 2022) on GitHub: individuals taking GitHub repositories that have a license and reposting it without the license.

[40] Context may change how relevant provenance is. For example, we could use MNIST to train a generative model, and we know the provenance of MNIST; however, we would not necessarily understand the implications of MNIST in generations in the same way as we understand its implications for classification.

[41] Northcutt et al. (2021) investigated mislabeled images in MNIST. Anecdotally, Yann LeCun has been overheard claiming he knows every mislabeled image in MNIST.

[42] Of course, datasets can also become stale if analysis takes too long. Just as choices are everywhere, so are trade-offs (Cooper et al., 2021).

Despite this, there has been significant push within the ML community for dataset creators to document their datasets before releasing them. One common recommendation is to create a datasheet that organizes information about how the data was collected, the motivation behind it, any preprocessing that was done, and future maintenance plans (Gebru et al., 2021).[43] As an example, this is The Pile's datasheet (Biderman et al., 2022). However, even an extensive datasheet like The Pile's still answers only a tiny fraction of the questions you could ask about what data it contains, why it was included, and how it was collected.

2.4 Conclusion & Next: Copyright and Training Data

By now, it's hopefully clear just how many choices dataset creators and curators make to produce and maintain datasets. To describe some of these choices, we discussed early dataset history, and showed how text and image generation systems once didn't rely on training data in the contemporary sense. For example, after the move from more classical AI rule-based techniques, text datasets involved carefully curated, hand-annotated data. Today, in contrast, the massive datasets collected to train generative AI models are scraped from the web, and are far too large to manually examine completely.

The size of these datasets has presented a slew of new types of choices when creating datasets, such as what (and how) to include and exclude different types of data. The lack of consensus on what to include in datasets is a reflection of the lack of societal consensus on what we want the capabilities of generative AI to be, in addition to the technical limitations we currently face. Today's datasets are shaped by today's influences: present model sizes, availability of data and compute, open-ended goals (and sometimes, a lack of desire to specify a specific goal), business incentives, and user desires. Ultimately, the way tomorrow's datasets are collected and curated will depend on factors that model the influences of users, governments, and businesses.

One particular concern from today's dataset creation practices is that scraped datasets may contain data with many different owners. In contrast to prior practices of curating public domain and licensed data, the choice to use scraped datasets with unclear provenance and documentation can raise copyright issues. Next in this series on generative AI, we'll discuss what sorts of copyrightable works could be included in the training data, why they may have ended up there, and whether or not that is permissible. Additionally, we'll discuss how different media (text, image, video, music, etc.) might require different treatment.

[43] Many datasets available on HuggingFace (a popular open-source model and dataset repository) now have datasheets attached to them.

Bibliography

Accessible Publishing. Guide to Image Descriptions, 2023. URL https://www.accessiblepublishing.ca/a-guide-to-image-description/.

Esther Ajao. The effect of reddit's decision to charge for data use. *TechTarget*, April 2023. URL https://www.techtarget.com/searchenterpriseai/news/365535524/The-effect-of-Reddits-decision-to-charge-for-data-use?Offer=abMeterCharCount_var1.

Maria Antoniak and David Mimno. Bad Seeds: Evaluating Lexical Methods for Bias Measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.148. URL https://aclanthology.org/2021.acl-long.148.

Authors Guild v. Google, 2015. URL https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022. URL https://www.anthropic.com/index/training-a-helpful-and-harmless-assistant-with-reinforcement-learning-from-human-feedback.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the Pile, 2022.

Geoffery C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA, USA, 2000. ISBN 0262522950.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What Does It Mean for a Language Model to Preserve Privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534642. URL https://doi.org/10.1145/3531146.3534642.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium*

(USENIX Security 21), pages 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. Technical report, Google, 2013. URL http://arxiv.org/abs/1312.3005.

Clarkson, et al. v. OpenAI, et al., 2023. URL https://clarksonlawfirm.com/wp-content/uploads/2023/06/0001.-2023.06.28-OpenAI-Complaint.pdf.

William W. Cohen. Enron Email Dataset. Technical report, Carnegie Mellon University, 2015. URL https://www.cs.cmu.edu/~./enron/.

Computer Graphics - The University of Utah, 1975. URL https://graphics.cs.utah.edu/teapot/. The Utah Teapot.

A. Feder Cooper, Karen Levy, and Christopher De Sa. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416.3483289. URL https://doi.org/10.1145/3465416.3483289.

A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 864–876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533150. URL https://doi.org/10.1145/3531146.3533150.

Nick Couldry and Andreas Hepp. *The Mediated Construction of Reality.* Polity Press, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. URL https://ieeexplore.ieee.org/document/5206848.

Maria Deutscher. Getty Images sues Stability AI for copyright and trademark infringement. *SiliconANGLE*, 2023. URL https://siliconangle.com/2023/02/06/getty-images-sues-stability-ai-copyright-trademark-infringement/.

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, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

William Brett Fishburne. Checkmates for Four Pieces, 2003. URL https://www.gutenberg.org/ebooks/4656.

W. Nelson Francis and Henry Kucera. Brown Corpus Manual. Technical report, Brown University, July 1979. URL http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for Datasets, 2021.

GitHub Copilot litigation, 2022. URL https://githubcopilotlitigation.com/.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Google. Google Ngram Viewer, 2012. URL http://books.google.com/ngrams/datasets.

Google Operating System. Google Switches to Its Own Translation System, 2007. URL http://googlesystem.blogspot.com/2007/10/google-translate-switches-to-googles.html.

David Graff and Christopher Cieri. English Gigaword. Technical report, Linguistic Data Consortium, Philadelphia, 2003. URL https://catalog.ldc.upenn.edu/LDC2003T05.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, 2022. URL https://arxiv.org/abs/2203.15556.

Jigsaw. Better Discussions with Imperfect Machine Learning Models, 2017. URL https://medium.com/jigsaw/better-discussions-with-imperfect-models-91558235d442.

Sheldon Klein, John F. Aeschlimann, David F. Balsiger, Claudine Converse, Steven L. Court, Mark Foster, Robin Lao, John D. Oakley, and Joel Smith. Automatic Novel Writing: A Status Report. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1973. URL https://minds.wisconsin.edu/handle/1793/57816.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, September 13-15 2005. URL https://www.statmt.org/europarl/.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

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, 2023. URL https://huggingface.co/bigscience/bloom#training-data.

Quoc V. Le and Mike Schuster. A Neural Network for Machine Translation, at Production Scale. Technical report, Google Research, September 2016. URL https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database, 1999. URL https://www.lri.fr/~marc/Master2/MNIST_doc.pdf.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014. ISBN 978-3-319-10602-1. URL https://cocodataset.org/#home.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. URL https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity, 2023.

Alexandra Sasha Luccioni and David Rolnick. Bugs in the Data: How ImageNet Misrepresents Biodiversity, 2022. URL https://arxiv.org/abs/2208.11695.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. Treebank-3. Technical report, Linguistic Data Consortium, Philadelphia, 1999. URL https://catalog.ldc.upenn.edu/LDC99T42.

Abubakar Mohammed. GitHub Copilot AI Is Generating And Giving Out Functional API Keys. *FOSSBYTES*, July 2021. URL https://fossbytes.com/github-copilot-generating-functional-api-keys/.

Netflix. Netflix Prize data, 2009. URL https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018.

Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, December 2008. URL https://www.robots.ox.ac.uk/~vgg/data/flowers/102/.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. *J. Artif. Int. Res.*, 70:1373–1411, may 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12125. URL https://doi.org/10.1613/jair.1.12125.

OpenAI. DALL-E 2, 2023. URL https://openai.com/dall-e-2.

Paul Tremblay and Mona Awad v. OpenAI, et al., 2023. URL https://torrentfreak.com/images/authors-vs-openai.pdf.

Project Gutenberg, 2023. URL https://www.gutenberg.org.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling Language Models: Methods, Analysis & Insights from Training Gopher, 2022. URL https://arxiv.org/abs/2112.11446.

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. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/reed16.html.

Reuters Staff. What does Twitter 'rate limit exceeded' mean for users? *Reuters*, July 2023. URL https://www.reuters.com/technology/what-does-twitter-rate-limit-exceeded-mean-users-2023-07-03/.

Eli Rosenberg. Facebook censored a post for 'hate speech.' It was the Declaration of Independence. *The Washington Post*, 2018. URL https://www.washingtonpost.com/news/the-intersect/wp/2018/07/05/facebook-censored-a-post-for-hate-speech-it-was-the-declaration-of-independence/.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage?, 2023. URL https://arxiv.org/abs/2304.15004.

Christoph Schuhman. LAION-AESTHETICS, August 2022. URL https://laion.ai/blog/laion-aesthetics/.

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

Stability AI. Stable diffusion public release, August 2022. URL https://stability.ai/blog/stable-diffusion-public-release.

Stable Diffusion litigation, 2023. URL https://stablediffusionlitigation.com/.

Yi Tay and Mostafa Dehghani. UL2 20B: An Open Source Unified Language Learner. Technical report, Google Research, October 2022. URL https://ai.googleblog.com/2022/10/ul2-20b-open-source-unified-language.html.

Chance Townsend. Twitter's copyright system seemingly broken as full-length movies are posted on platform. *Mashable*, 2022. URL https://mashable.com/article/twitter-copyright-full-movies.

Scott R. Turner. *MINSTREL: A computer model of creativity and storytelling.* PhD thesis, University of California, Los Angeles, 1993. URL https://www.proquest.com/docview/304049508.

Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss. Does GPT-2 Know Your Phone Number? *Berkely Artificial Intelligence Research*, December 2020. URL https://bair.berkeley.edu/blog/2020/12/20/lmmem/.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Joseph Weizenbaum. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM*, 9(1):36–45, jan 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL https://doi.org/10.1145/365153.365168.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. URL https://authors.library.caltech.edu/27452/1/CUB_200_2011.pdf. CNS-TR-2010-001.

Wheaton v. Peters, 1834. URL https://supreme.justia.com/cases/federal/us/33/591/.

Wikipedia. Licence laundering, 2022. URL https://en.wikipedia.org/wiki/Licence_laundering.

Wikipedia. Procedural texture, 2023. URL https://en.wikipedia.org/wiki/Procedural_texture.

Wallace Witkowski. Reddit founder wants to charge Big Tech for scraped data used to train AIs: report . *MarketWatch*, April 2023. URL https://www.marketwatch.com/story/reddit-founder-wants-to-charge-big-tech-for-scraped-data-used-to-train-ais-report-6f407265.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against Neural Fake News. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. URL https://rowanzellers.com/grover/.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM*, 64(3):107–115, feb 2021. ISSN 0001-0782. URL https://doi.org/10.1145/3446776.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, Los Alamitos, CA, USA, December 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.11. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.11.