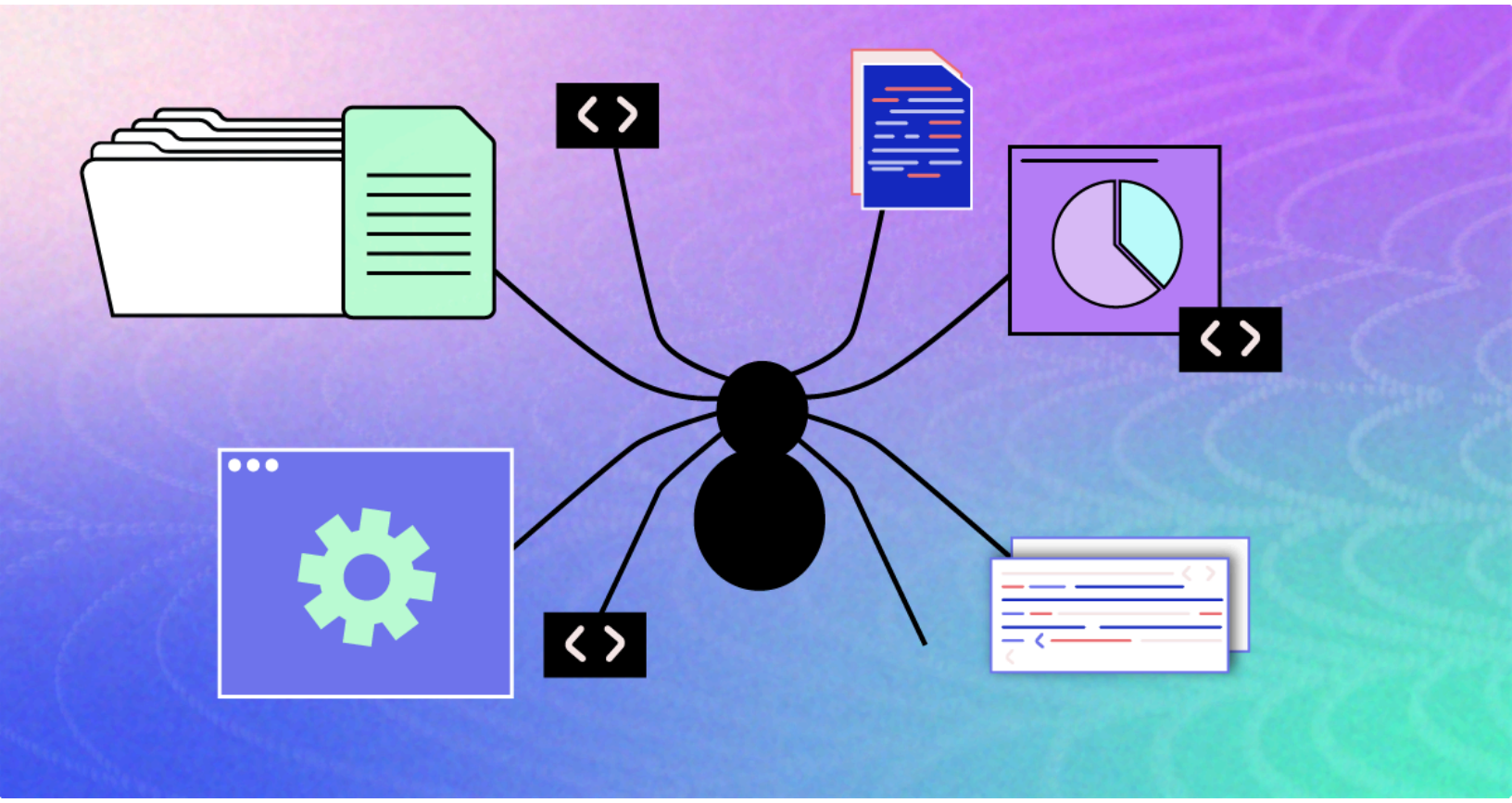
moz://a

Training Data for the Price of a Sandwich[1]

Common Crawl's Impact on Generative AI



February 2024

By Stefan Baack

Mozilla Insights

[1] Title inspired by "Analyzing the Web For the Price of a Sandwich," an article about Common Crawl.

About Mozilla

Mozilla's mission is to ensure the internet is a global public resource, open and accessible to all. An internet that truly puts people first, where individuals can shape their own experience and are powerful, safe, and independent.

Founded as a community open source project in 1998, Mozilla currently consists of two organizations: the nonprofit Mozilla Foundation, which leads our movement building work; and its wholly owned subsidiary, the Mozilla Corporation, which leads our market-based work, including the development of the Firefox web browser. The two organizations work in close concert with each other and a global community of tens of thousands of volunteers under a single banner: Mozilla.

Acknowledgements

I would like to thank the following reviewers for their time and invaluable feedback. In alphabetical order: J. Bob Alotta, Ziyaad Bhorat, Stella Biderman, Abeba Birhane, Maximilian Gahntz, Lisa Gutermuth, Alex Hanna, Stephen Hood, Bernard Koch, Solana Larsen, Crystal Lee, EM Lewis-Jong, Eeva Moore, Kasia Odrozek, Will Orr, Julian Posada, Kenrya Rankin, Victor Storchan, and Apryl Williams. I would also like to thank Common Crawl for the interviews and their comments on the report prior to the publication.

Table of contents

Executive Summary

Common Crawl is a small nonprofit organization that has created a massive (9.5-plus petabytes), freely available archive of web crawl data dating back to 2008. This data has been valuable for many researchers, but since 2020 when OpenAI published GPT-3, the large language model (LLM) that still powers the free version of ChatGPT, Common Crawl has become one of the most important sources of training data for generative AI. However, Common Crawl's role in generative AI has received relatively little attention so far. In this report, we look at Common Crawl in-depth and ask what its popularity means for trustworthy AI.

Common Crawl primarily wants to level the playing field for technology development, not just provide data for AI training. It was founded in 2007 with the intention to mimic the way Google crawled the web for its search engine. Common Crawl's goal is to make both the kinds and the amounts of data that usually only big tech companies like Google have access to available to researchers and smaller businesses.

Common Crawl's mission as an organization does not easily align with the needs of trustworthy AI development. Its guiding principle is that less curation of the provided data enables more research and innovation by downstream users. Common Crawl therefore deliberately does not remove hate speech, for example, because it wants its data to be useful for researchers studying hate speech. However, such data is undesirable when training LLMs because it might lead to harmful outputs by the resulting models.

Common Crawl does not contain the "entire web," nor a representative sample of it. Despite its size, there are important limitations on how much of the web is covered. The crawling process is almost entirely automated to prioritize pages on domains that are frequently linked to, which makes domains related to digitally marginalized communities less likely to be included. The language and regional coverage is strongly skewed toward English content. Moreover, a growing number of relevant domains like Facebook and the New York Times block Common Crawl from crawling most (or all) of their pages.

When used as a source for AI training, Common Crawl should be used with care, but such care is often lacking. Due to Common Crawl's deliberate lack of curation, AI builders do not use it directly as training data for their models. Instead, builders choose from a variety of filtered Common Crawl versions to train their LLMs. However, there is a lack of reflection among AI builders about the limitations and biases of Common Crawl's archive. Popular Common Crawl versions are especially problematic when used to train LLMs for end-user products because the filtering techniques used to

create them are simplistic and often focused on removing pornography or boilerplate text like the names of navigational menu items, leaving lots of other types of problematic content untouched.

Common Crawl and AI builders have a shared responsibility for making generative AI more trustworthy. While Common Crawl was never primarily about providing AI training data, it now positions itself as an important building block for LLM development. However, it continues to provide a source that AI builders need to filter before model training. Both groups can help make generative AI more trustworthy in their own ways. Common Crawl should better highlight the limitations and biases of its data and be more transparent and inclusive about its governance. It could also enforce more transparency around generative AI by requiring AI builders to attribute their usage of Common Crawl. AI builders should put more effort into filtering out more types of problematic content and try to better take into account the various cultural contexts in which their generative AI products are deployed. There is also a need for industry standards and best practices for end-user products to reduce potential harms when using Common Crawl or similar sources for training data. In addition, AI builders should create or support dedicated intermediaries tasked with filtering Common Crawl in transparent and accountable ways that are continuously updated. Long term, there should be less reliance on sources like Common Crawl and a bigger emphasis on training generative AI on datasets created and curated by people in equitable and transparent ways.

Introduction

"Often it is claimed that Common Crawl contains the entire web, but that's absolutely not true. Based on what I know about how many URLs exist, it's very, very small." (Interview, main crawl engineer at Common Crawl)

Common Crawl (henceforth also referred to as CC) is an organization that has been essential to the technological advancements of generative AI, but is largely unknown to the broader public. This California nonprofit with only a handful of employees has crawled billions of web pages since 2008 and it makes this data available without charge via Amazon Web Services (AWS). Because of the enormous size and diversity (in terms of sources and formats) of the data, it has been pivotal as a source for training data for many AI builders. Generative AI in its current form would probably not be possible without Common Crawl, given that the vast majority of data used to train the original model behind OpenAI's ChatGPT, the generative AI product that set off the current hype, came from it (Brown et al. 2020). The same is true for many models published since then.

Although pivotal, Common Crawl has so far received relatively little attention for its contribution to generative AI. Its size and prevalent usage across many models is inadvertently hinted at in (dubious) claims that generative AI products have been trained on (nearly) the "entire internet" (see for example McKinsey 2023; or Morrison 2023). Common Crawl has recently come under fire for potentially including unauthorized copyrighted material in the corpus it makes freely available. The New York Times sued OpenAI and Microsoft because their models were allegedly trained on its content. One piece of evidence provided in the lawsuit is that Common Crawl contained a substantial amount of content from NYTimes.com when OpenAI launched ChatGPT (the Times has since pushed Common Crawl to remove this content from its archives). Like other content creators, the Times is concerned that AI models trained on Common Crawl will be able to mimic its work without providing compensation. More and more platforms, online communities, and news media want to block or charge money for access to their data. An increasingly important component of those efforts is blocking or misdirecting web crawlers, including Common Crawl's.

However, those claims and controversies suggest a lack of understanding about Common Crawl and the data it provides. First, Common Crawl has never solely been about providing AI training data. It was founded in 2007, long before the advent of generative AI and even before terms like "big data" became popular, and for most of its existence the majority of its users were researchers across various fields. Second, while the data it provides is massive (more than 9.5 petabytes), it is far from representing the

“entire web,” which Common Crawl emphasizes too (see below). Uncritically assuming that generative AI models have been trained on the entire web by simply including a lot of data from Common Crawl washes over the shortcomings and biases inherent in this data. Third, it is worth noting that Common Crawl does not contain complete replicas of specific web domains (and their potentially copyrighted materials).

This lack of understanding is problematic because what data AI builders choose to train their models on is important, as it has downstream effects for how these models behave and what they are useful for. Considering the impact of generative AI on societies worldwide, this report sheds light on the implications of Common Crawl’s popularity by exploring the values that guide Common Crawl itself in-depth.

A lot of critical research and journalistic reporting about AI training data looks at its contents, and some have also highlighted what is included in Common Crawl’s corpus specifically (see for example Luccioni and Viviano 2021; Schaul, Chen, and Tiku 2023). This work is important for assessing fairness and bias. However, in this report, we take a different approach to assess how the values and practices of Common Crawl affect AI models trained on it by studying its data as *infrastructure*. Common Crawl is never used directly to train generative AI because it contains too much data that is considered undesirable for AI model training. Instead, AI builders filter Common Crawl before the training; there are a handful of filtered Common Crawl versions that are used frequently. Common Crawl therefore has an infrastructural role within generative AI research and development in the sense that it provides a basis from which AI builders create training datasets. Infrastructures shape the practices and sense-making of the people who use them. As media studies scholar Luke Munn argues:

> *One of the things that make infrastructures so powerful is that they model their own ideals. They privilege certain logics and then operationalize them. And in this sense… they both register wider societal values and establish blueprints for how they should be carried out. (Munn 2022)*

Given Common Crawl’s role as a foundational building block for many generative AI models, studying its influence as infrastructure allows us to evaluate the stated values and intentions that guide the creation and continued maintenance of its data mean for AI models trained on it. To explore this, this report is structured as follows:

1. We start with a discussion of our research methodology and its strengths and limitations.
2. Then we explain how Common Crawl’s data is used to train generative AI and how frequently AI builders have relied on it.

3. Next, we examine the values and practices of Common Crawl itself, and how its data is collected.
4. Then we take a closer look at how AI builders filter Common Crawl's data before training their models with it.
5. Finally, we discuss the implications of Common Crawl's popularity for trustworthy AI and ask what needs to change to build more trustworthy generative AI products.

Methods and limitations

To study the influence of Common Crawl on generative AI, we followed Denton et al. (2020) and looked at the histories, values, and norms embedded in its datasets. This means examining the motivations, assumptions, and values of the people at Common Crawl who shape the purpose and design of the crawl. To achieve this, our approach included the collection of quantitative data about Common Crawl's prevalence as a data source for generative AI, but primarily relied on the qualitative analysis of various sources (see a detailed list below).

To provide evidence about Common Crawl's importance for generative AI, we collected information about generative AI models released between 2019 and October 2023. Given the speed at which new models are released, creating a comprehensive list of every single model published within this time frame turned out to be very difficult, however. To make our work more feasible, we decided early on to limit ourselves to text generators used for applications like ChatGPT, not image generators and multimodal models. We also limited our collection to pre-trained, not adapted models (this distinction is explained in the next section). In addition to the general difficulties of compiling a comprehensive list, our collection also suffers from a lack of transparency, especially when it comes to some of the more popular models, such as OpenAI's GPT-4 and Meta AI's Llama v2, that do not provide any information about their training data. Despite these shortcomings, we're confident that our data aptly reflects the industry's reliance on Common Crawl.

Our primary sources for collecting text generator models were two academic papers that were continuously updated throughout 2023: Yang et al. (2023) and Zhao et al. (2023). Both are meant as practical introductions to large language models (LLMs) and provide overviews of published models. They were chosen based on cross-referencing with other academic and non-academic collections (for example Kim 2023). For each model, we reviewed its accompanying research paper and recorded a) whether Common Crawl's data had been used and b) what specific version of Common Crawl was utilized. Our full list is available here.

To gain a deep understanding of Common Crawl itself, we relied on various qualitative data sources. First, we conducted semi-structured interviews, each approximately 40 minutes long, with two pivotal members of Common Crawl's staff: its current director and its main crawl engineer.[2] Both interviewees were asked how Common Crawl reflects the web, how the crawl is conducted, how the data is processed, and about

[2] The interview with the crawl engineer was conducted in German, all quotes from it were translated into English by the author.

Common Crawl's relationship with builders of generative AI. Specifically tailored to their roles, the director was asked about the mission and purpose of the organization, and the crawl engineer about the intricacies of the crawl itself and the organization's history (given his tenure was longer than the director's, who was appointed a few months prior to the interview).

Common Crawl only had three employees when we collected our data. We therefore collected additional materials to increase our understanding of Common Crawl's impact on generative AI:

- Eight discussion threads from Common Crawl's public mailing list: We first examined all 376 discussion threads from January 2020 (the year OpenAI published GPT-3) to October 2023 and selected those related to the representativeness of the data for the entire web, the organization's relationship with AI builders, its mission, and explanations of how the crawl works. Interesting for our analysis were individual posts by Common Crawl staffers, not the entire discussion threads. Following the initial analysis, we used keywords to search the archives going back to 2011 with 1157 discussion threads. Those keywords were informed by our interviews and our manual review of more recent discussions on the mailing list: [LLM], [LLMs], ["large language model"], [OpenAI], ["Open AI"], [C4], ["Colossal Clean Crawled Corpus"], [AI], [NLP], [quality], [blekko].[3]
- Common Crawl's website: We reviewed all pages of the website. We also included three relevant posts from the blog (which dates back to 2011).
- Presentation slides created by the main crawl engineer: Prior to our interview, the main crawl engineer shared slides he prepared for builders in the natural language processing (NLP) field. See Nagel (2023).
- Three published interviews with Common Crawl's founder and Chairman Gil Elbaz: We were particularly interested in his motivations and vision guiding the foundation of Common Crawl in 2007. We searched Google with the following keywords: ["Gil Elbaz" + "Common Crawl"], ["Gil Elbaz" + "Factual"], ["Gil Elbaz" + "interview"], ["Gil Elbaz" + "Google"]. We picked interviews by Cremades (2019), Rogers (2014), and Zaino (2012) as most relevant for filling our knowledge gaps.
- Two articles highlighting perspectives from past leadership: We included two interviews with two different former directors of Common Crawl, Lisa Green (Owens 2014) and Sara Crouse (Leetaru 2017), that provided relevant information about Common Crawl's historical trajectory and self-perception.

[3] Square brackets are used to indicate search terms, meaning inside the brackets are the exact searches used. "blekko" is the name of a search engine startup that historically played an important role for Common Crawl.

This qualitative data was analyzed using qualitative coding with the Taguette QDA software (Rampin and Rampin 2021). We used inductive qualitative coding following Braun and Clarke's (2012) thematic analysis to guide the discovery of recurrent themes across our various data sources. These themes provided us with valuable insights related to our primary interest in the downstream effects of Common Crawl's prominence for generative AI.

As Common Crawl offers its data in individual crawls that have to be downloaded separately (more on this below), our analysis is limited to understanding how the crawl data that most LLM builders used to train their models since 2020 was collected (which roughly covers 2017 to 2023). While we do look into Common Crawl's history to the degree that it is still relevant for more recent crawls, we did not conduct a more in-depth analysis into how the crawl has evolved over time, since Common Crawl's founding in 2007. If AI builders include a significant amount of crawl data from before 2017, this history is relevant, as both the frequency with which crawls were conducted and how Common Crawl discovered new URLs to crawl differed significantly (Nagel 2023). A more in-depth analysis of Common Crawl's history could have involved interviewing former Common Crawl employees, including founder Gil Elbaz.

Finally, because our focus is on how Common Crawl works and how it perceives its own role within the AI ecosystem, we did not investigate in-depth how AI builders use and think about Common Crawl's data, for example by conducting additional interviews with relevant builders. While we do look into how AI builders filter Common Crawl, we deemed a review of research papers published by builders sufficient. Our work forms a useful basis for future research in this direction and complements existing work like Orr and Crawford's (2023) study.

Common Crawl's role in generative AI

To clarify Common Crawl's role in generative AI, it is important to understand that there are roughly two phases in the production process of generative AI products: the pre-training phase and the fine-tuning phase.

The pre-training phase typically refers to training an LLM that is good at predicting the next token in a sequence, which can be the next word (or part of a word) in a sentence, the next pixel in an image, or something else. While there are generative AI models that do not rely on LLMs, most of them do (especially text generators) and we will focus on this type of generative AI for the purposes of this report. Training LLMs requires massive amounts of data, too massive to manually create, label, or review it. Therefore, pre-training relies on automated and scalable data creation and curation techniques, and on unsupervised learning (meaning the AI model is not given labeled data to guide how the information should be classified during training). The resulting LLM is essentially a massive token prediction machine that continues any given sequence (for instance, a prompt in a chatbot).

By themselves, pre-trained LLMs are not very useful to most people. Getting a pre-trained LLM to produce a desired output can require a lot of work because they do not reliably follow instructions, and without any adaptation they are very likely to (re)produce problematic, unverified, or biased content that may have been part of their pre-training data. Pre-trained LLMs can form the basis for many different applications, however, which is why they have also been called foundation models (Bommasani et al. 2021). As Microsoft's CTO Kevin Scott stated, he does not consider these models "products," but "infrastructure. They are building blocks that you use to make products, but they are not products themselves" (in Patel 2023). This is possible because LLMs can be modified after their initial pre-training.

The fine-tuning phase is about modifying LLMs so that they more reliably produce desired outputs. A lot of different fine-tuning techniques exist, and new ones keep emerging. One prominent approach is "reinforcement learning from human feedback" (RLHF), which was used by OpenAI to modify its LLM called GPT-3 to create ChatGPT (Ouyang et al. 2022). Among other things, RLHF involves data workers rating multiple responses by the LLM to the same prompt (from best to worst), following instructions from AI builders and writing ideal question and answer pairs for training. By optimizing their model to produce outputs rated highly by data workers, AI builders make LLMs act more predictably. Other, cheaper techniques to adapt LLMs that do not require hiring data workers are prompt engineering and Retrieval Augmented Generation (RAG).

In the following, we will focus on pre-training because this is where Common Crawl plays a crucial role.

Pre-training data curation for text generators

As mentioned above, LLM builders typically seek to maximize the quality, size, and diversity of their training data. This makes Common Crawl useful because it is a massive collection of mostly HTML code and text extracted from billions of URLs across the web, about 9.5 petabytes in total as of mid-2023 (Interview CC Director). Especially in natural language processing (NLP), Common Crawl has been used for almost a decade already (Hayes 2015), well before the recent boom of generative AI (one of the earliest examples is Pennington, Socher, and Manning 2014). However, with the invention of the transformer technology (Vaswani et al. 2017) that enabled a major leap in generative AI, the demand for greater volumes of training data increased. This made Common Crawl more popular. Early generative AI models actually relied on it even more than most newer ones, sometimes using nothing other than Common Crawl for the pre-training (Gao et al. 2020).

Exclusively relying on Common Crawl's data to train AI models has downsides, however. It contains large amounts of content that is undesirable for AI training: problematic content like hate speech and pornography (Luccioni and Viviano 2021), and low quality content like "boiler-plate text like menus, error messages, or duplicate text" (Raffel et al. 2019). To mitigate this, AI builders use filtered versions of Common Crawl for the training.

Since 2020, AI builders increasingly use a collection of various (relatively) small datasets, like timebound Wikipedia snapshots that they consider "high quality." Doing so was found to potentially "improve the general cross-domain knowledge and downstream generalization capabilities of the model" (Gao et al. 2020) Filtered versions of Common Crawl continue to be essential for many models, however, as AI builders use them to scale up the size of their pre-training data to ensure their LLMs have the desired performance. The overall proportion of Common Crawl often remains significant. For example, Common Crawl made up more than 80% of the tokens in OpenAI's GPT-3 (Brown et al. 2020).

A review of 47 LLMs for text generation published between 2019 and October 2023 shows that at least 64% of these models (30) used at least one filtered version of Common Crawl for their pre-training (Figure 1).

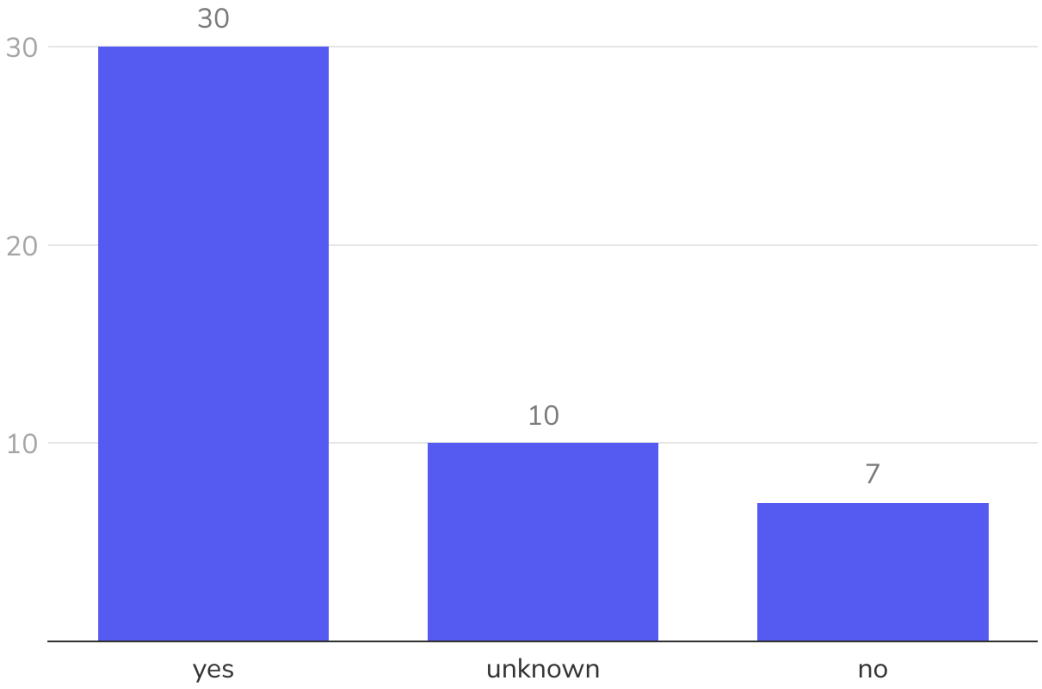


Figure 1: Count of 47 text generator LLMs published between 2019 and October 2023 using Common Crawl for their pre-training. "Unknown" refers to instances where AI builders did not disclose enough information about the pre-training data to determine whether Common Crawl was used.

Most AI builders do not filter Common Crawl themselves, but rely on versions published by others (Figure 2). Popular filtered versions are Alphabet's "Colossal Clean Crawled Corpus" ("C4," see Raffel et al. 2019) and "Pile-CC," a version of Common Crawl included in EleutherAI's LLM training dataset "The Pile" (Gao et al. 2020). Note that in some cases, AI builders use more than one filtered version. For example, Meta AI's Llama v1 used two because the authors claim that "using diverse pre-processed CommonCrawl [sic] datasets improves performance" (Touvron, Lavril, et al. 2023).

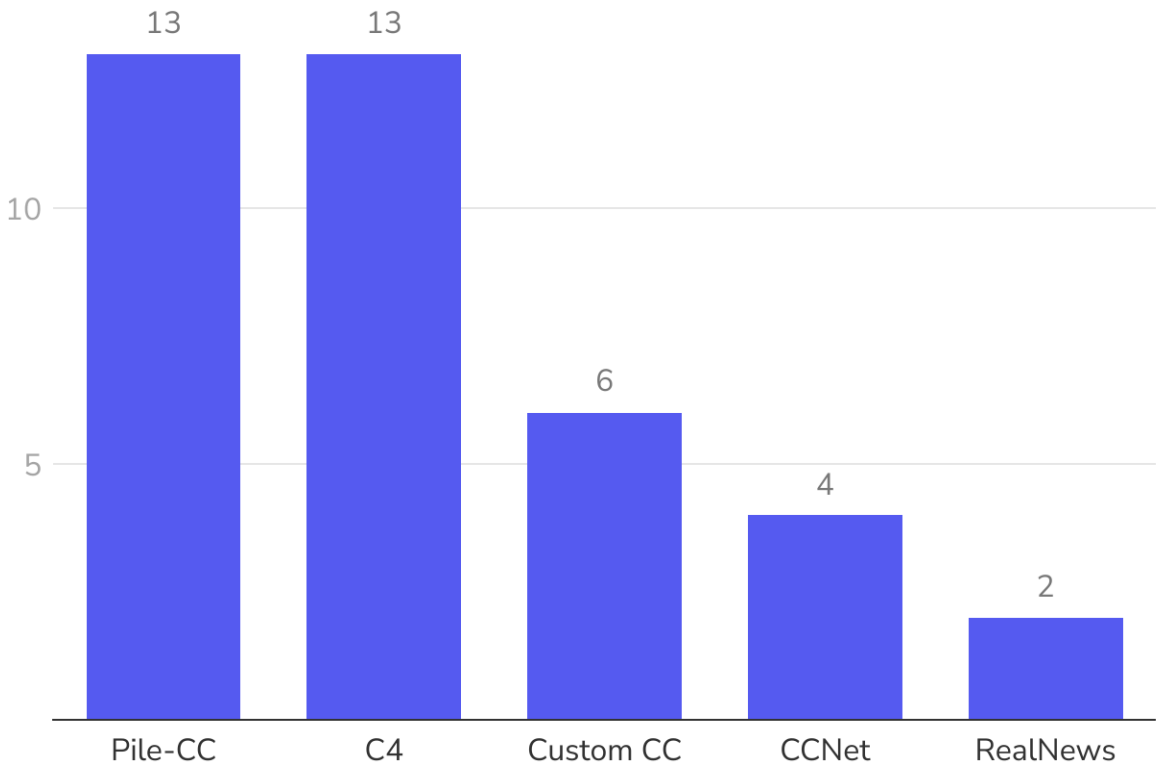


Figure 2: Five most used filtered versions of Common Crawl among 47 text generator LLMs published between 2019 and October 2023. The count shows how many LLMs used this version for their pre-training. "Custom CC" means that AI builders a) applied their own filtering, rather than relying on filtering techniques published elsewhere, and b) their exact filtering techniques were not used elsewhere.

The practice of scaling up pre-training data with filtered Common Crawl versions is so widespread that the BigScience workshop, a huge international undertaking from 2021 to 2022 to create an open LLM as an alternative to the offerings of leading AI companies, added a filtered version of Common Crawl to its training data because "not including it would constitute too much of a departure and risk invalidating comparisons [with previously released LLMs]" (Laurençon et al. 2023). The comment indicates just how much Common Crawl has shaped AI builders' expectations of their model's performance and behavior. To better understand Common Crawl's influence, we will now take a closer look at the organization and the data it provides.

Common Crawl's mission: Enabling others to work like Google

Common Crawl's mission and values as an organization stem from those of its founder, Gil Elbaz, who not only created the nonprofit organization in 2007, but has financed it and acted as its chairman ever since (Interview CC Director). Common Crawl has always been small in terms of staff, and Elbaz is the only individual who has been involved continuously without interruptions throughout its lifecycle.

Before Common Crawl, Elbaz was one of the co-founders of Applied Semantics in 1999, a company that gathered and categorized information on websites to serve contextual advertisements using a technology called AdSense. In 2003, Applied Semantics was acquired by Google (now Alphabet), and Elbaz joined Google and worked there until 2007. In later interviews, Elbaz described how his experience at Google led him to leave and start what he refers to as "neutral data companies":

> *It was amazing, but it also refined and continued to shape my world view [sic] that the data moat is an incredible advantage that Google has.... I became a little bit concerned that Google could become a monopoly of innovation. I felt like a world where many companies are bringing innovation forth, across the world, I felt like that ultimately is the world that I want to live in. So, I started to think about creating a neutral data company, a company that wants to democratize access to information to provide data to other companies.... That's what we ended up doing.* (Elbaz in Cremades 2019)

Elbaz is not referring to Common Crawl here, but to Factual, a for-profit company he founded shortly after Common Crawl in 2008. Factual (which merged with Foursquare in 2020) specialized in gathering and structuring location data, especially about local businesses, and sold access to this refined data to other companies, including Microsoft. While only referring to Factual, the vision of democratizing information access with "neutral data companies" Elbaz describes above is applicable to Common Crawl as well. Common Crawl's stated mission is to provide access to "high quality crawl data that was *previously only available to large search engine corporations*" to "small startups or even individuals" (Common Crawl Foundation, n.d. emphasis added). In other words, Common Crawl's purpose is to make web data available that otherwise only a Big Tech company would have access to.

The coexistence of Common Crawl as a nonprofit and Factual as a for-profit business sheds light on Elbaz's vision. In an article from 2012 about Common Crawl's role in the data tech ecosystem, he is quoted about the relationship as follows:

> *Elbaz expects that Common Crawl and efforts that build upon its data may help lay the foundation of things that will be valuable to applications over time, as well. He posits, for example, an end user trying to create a restaurant search app and the access she now can have to more resources to develop it, using Common Crawl directly for free, or looking up clean, structured Global Place data from Factual (for which the pre-existing open crawl is one of multiple resources) for a small fee.... "Each additional resource creates that much more productivity and manifests itself in better and better consumer applications," Elbaz says.* (Zaino 2012)

Common Crawl is meant to be a public resource, an infrastructural building block in Elbaz's vision that provides largely unrefined web crawl data to anyone for free. The current director also emphasized the infrastructural quality of Common Crawl that should help jump-start and uplift its users:

> *You know, why do you need Common Crawl? It's all out there on the web, you can just go get it yourself. But it's difficult to start and operate a web crawler, so if you're a researcher and you want to do some kind of study but need a billion pages before you can start, that's a lot of work and there are a lot of issues involved with that.* (Interview CC Director)

The unrefined nature of the data is essential to this notion. Staffers of Common Crawl frame this as a tradeoff. Offering the largely unrefined crawl data means that in most cases, this data cannot be used directly without further refinement. At the same time, the lack of curation is seen as enabling open-ended research and innovation, as the data is potentially useful to a greater number of use cases than a more curated crawl. As the director put it:

> *From a goal standpoint, I don't think we want to necessarily be curating the dataset because the pages we removed might be of value to downstream users. You might be looking for the prevalence of hate speech within a certain country...if you're the researcher trying to measure the prevalence, you want that material in there. So we kind of said it's sort of up to the downstream user to do content classification.* (Interview CC Director)

Factual exemplified a business made easier by this type of nonprofit data infrastructure; it was a data collection and cleaning warehouse that sold refined data to others.

Common Crawl's data: Machine scale analysis

The "Google-style," data-driven innovation Elbaz wanted to support is directly reflected in how Common Crawl collects and provides its data. The process is designed to support what former director Lisa Green called "machine scale analysis" (in Owens 2014): automated, large-scale analysis of web data across web domains (Interview CC Director). Common Crawl targets "programmers, data scientists, researchers working with web data… the page captures are not really useful for 'end user' [sic] browsing the archives. Instead we try to support 'data users'" (Nagel 2019). In short: Common Crawl's data is not meant to support qualitative research or more in-depth analysis of individual domains. While both might be possible with some caveats, Common Crawl aims to support the breadth and automation of web analysis rather than its depth.

In practice, this means Common Crawl strives to collect enough data to enable large, cross-domain analysis, while staying within the bounds of the US fair use doctrine for copyrighted materials. Among other things, this means that Common Crawl only collects the HTML code of the crawled pages in most cases, no CSS styling, and in most cases no images and other media are retained (Leetaru 2017). It also never collects full copies of domains. There are no complete copies of Wikipedia in Common Crawl for example. Instead, URLs from domains are sampled and these samples vary in size depending on the domain's centrality score (more on this below). In addition, there is a fixed upper limit of how many URLs are included in a single crawl (Interview CC Director). The collected data is provided in three different formats; most relevant for AI builders are WARC files containing the HTML code of the pages with some meta data, and WET files containing the extracted plaintext from the HTML code.

In other words, working with Common Crawl's data means working exclusively — or primarily, depending on how the data is filtered — with text spread across billions of URLs sampled from web domains. While this has always been true, Common Crawl is not a homogeneous, singular dataset. Instead, the data is offered as individual crawls of varying size that have to be downloaded separately. When including Common Crawl in their pre-training data, AI builders combine multiple crawls, typically starting from the latest crawl available during the data collection and combining that with older crawls, often stretching back years of crawling data. However, modern models usually stay within a time frame that Nagel (2023) describes as Common Crawl's latest phase of data collection starting in 2017, where it relies on distinct crawler implementations and different "approaches to find and sample (prioritize) seeds and URLs." In this time frame, crawls are published monthly with three to five billion crawled URLs each.

Note that Common Crawl provides two types of crawls, offered separately: the main crawls, and a crawl specifically for news that is updated more frequently than the main one. We will focus on the main crawl because the news crawl is rarely used by AI builders to train their LLMs (only once in our sample of 47 text generation models).

How Common Crawl decides which URLs to crawl

When Common Crawl's current main crawl engineer joined the organization in 2016, his primary task was to make Common Crawl less dependent on external seed donations, which were its most important source for finding new URLs to crawl between 2013 and 2016 (Interview CC Crawl Engineer). Seed donations are annotated lists of URLs shared (donated) by other organizations. One of the most important seed donors was the search engine startup blekko, which was acquired by IBM in 2015. However, seed donations dwindled over time, and Common Crawl was interested in automating the discovery of new URLs which were considered of good quality. The classic way of automatically discovering new URLs — following links on crawled pages — tends to include a lot of low-quality web data, not least because it is prone to link spam.

Inspired by an independent nonprofit search engine project called Common Search (discontinued in 2018), Common Crawl developed an approach that builds on ranking web domains by calculating their harmonic centrality (Interview CC Crawl Engineer). Harmonic centrality measures the importance of a node in a network based on the distance this node has to all the other nodes, with shorter distances contributing more to the score than longer ones. In other words, the more often a domain is directly or indirectly linked to, the higher its harmonic centrality score, with more direct links contributing more. Harmonic centrality essentially captures how accessible a domain is in the sense that it can be accessed via links on other pages. Compared to other centrality measures, it is a flat score that treats all domains the same. By contrast, Google's PageRank is based on eigenvector centrality, where the importance of a node is determined by how many other important nodes link to it. Common Crawl opted for harmonic centrality because it is deemed better at avoiding spam than other centrality measures (Interview CC Crawl Engineer). The scores are not only used to decide which domains to crawl, but also how many URLs from those domains to include (the higher the domain's score, the more URLs). To decide which URLs to sample from a particular domain, the domain's score is projected to all of its URLs and the highest ranking ones are included (Interview CC Crawl Engineer).

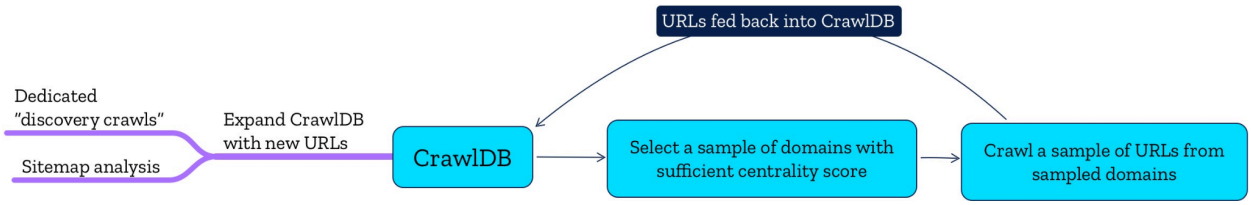


Figure 3: A simplified overview of Common Crawl's crawling process.

As illustrated in Figure 3, each crawl starts from a database called CrawlDB that contained about 25 billion URLs in August 2023 (Interview CC Crawl Engineer). For each URL, it records the score, a timestamp from when the URL was last fetched, whether it was crawled successfully, and more (Nagel 2022). The CrawlDB is regularly expanded by including URLs from previous main crawls, by conducting separate smaller crawls specifically designed for finding new URLs, and by analyzing sitemaps (Nagel 2022). To decide which URLs to pick from the CrawlDB for the next main crawl, Common Crawl combines the harmonic centrality score with additional score points added or subtracted depending on when the URL was last fetched successfully (with the goal of including more URLs that have not been crawled before or that haven't been included for some time). The score threshold necessary for a domain to be included is adjusted flexibly to reach the set maximum of URLs for one crawl. The selection also takes into account how many URLs per domain can be fetched depending on the domain's harmonic centrality score (Nagel 2022).

How Common Crawl defines quality data

Common Crawl staffers care about the quality of their data despite their emphasis on not curating it to enable open-ended innovation and research for downstream users. Indeed, quality was a frequent theme in our interviews and in Common Crawl's mailing list discussions. Quality is discussed in two ways: how Common Crawl's data represents the web as a whole and the quality of the URLs included in the crawls. In both of these discussions, Common Crawl draws on statistical notions of quality: the measurable diversity of languages or regions and a measurable notion of importance (harmonic centrality). This reliance on measurable indicators rather than qualitative assessments of the contents of the crawled data is compatible with Common Crawl's light approach to curation.

Regarding Common Crawl's representativeness of the web as a whole, there are important limitations for how much of the web is covered:

- Due to the methodology described above for picking URLs from the CrawlDB, about 50% of each crawl is made up of URLs that have been fetched in previous crawls (Interview CC Crawl Engineer). Because some top domains always have a very high harmonic centrality score (like Wikipedia) they are always included, while those with lower and varying scores may or may not be.
- Since 2017, crawls have been published monthly, with each crawl capturing between three and five billion URLs. Common Crawl's captures are discontinuous; what happens in between two different crawls is not captured in hindsight.
- Common Crawl respects robot.txt, a method used by web domain administrators to tell web crawlers which parts of the domain they are allowed to visit. Several large domains use robot.txt to block Common Crawl's crawler from visiting most or all of their content, like The New York Times and Facebook. The number of domains blocking Common Crawl is likely to increase in the wake of "data revolts" (Frenkel and Thompson 2023), where content creators aim to stop their data from being used to train LLMs.
- Common Crawl's technical infrastructure is based in the United States, which skews crawls toward English content (for example, pages providing multiple languages may serve the English version, Interview CC Crawl Engineer).

Given those limitations, staffers at Common Crawl strongly reject claims that their data represents the "entire" web. As the main crawl engineer put it:

> *That's something I try to explain to everyone: Often it is claimed that Common Crawl contains the entire web, but that's absolutely not true. Based on what I know about how many URLs exist, it's very, very small. I think that's really important.* (Interview CC Crawl Engineer)

In addition, staffers express uncertainty about *how* their data reflects the web as a whole. While the main crawl engineer is confident that the data is "reasonably representative" of the web in the sense that he thinks it contains a lot of "relevant data," he emphasizes that the problem is not knowing the size of the web as a whole: "Every year I keep losing faith in my understanding of the web. I have the impression that I know less and less about it" (Interview CC Crawl Engineer). The director similarly explains that "the web is practically infinite.... I don't have a good idea of what the comprehensiveness [of our data] is and I think even defining what you mean by comprehensiveness is a task in itself" (Interview CC Director).

The director calls Common Crawl's data an "academic sampling of the web" (Interview CC Director). Common Crawl's data indeed consists of samples of samples. Each crawl is made up of URLs sampled from web domains, and those domains are themselves sampled from the CrawlDB. While the CrawlDB itself is large, it is not designed to capture as much of the web as possible, as it does not add every URL Common Crawl finds, but instead relies on harmonic centrality scoring to moderate what is and is not included (Interview CC Crawl Engineer).

The decision to restrain the total sum of URLs to fetch per crawl is not only precipitated by limited resources, but also by a decision to balance the size and quality of the content that is included:

> *There is also a trade-off between how big you make the crawl and the quality of the crawl. When we stop at three billion, we think we're getting a pretty good sampling. If we were to 10X that, we believe the quality would drop because we're going into lower ranked pages and so you're going to get more junk and then people using it are going to have to filter out more stuff.* (Interview CC Director)

When discussing the quality of the data contained in its crawls, staffers of Common Crawl frequently discuss spam. In its early years, Common Crawl's stance toward spam was consistent with its foundational idea that less curation of the data increases the potential for open-ended research and innovation by downstream users. In a blog post from 2011, the organization argued that "it is not clear whether we *should* really remove spammy content from the crawl. For example, individuals who want to build a spam filter need access to a crawl with spam" (Common Crawl Foundation 2011). Today, avoiding and removing spam is a priority. Not only because spam is considered low-quality content, but because of "crawler traps," like domain parks designed to keep the crawler inside a web of connected spam pages. As the director explains, if the crawler wanders into such a trap "it can get stuck. And then you look at the crawler and 80% of it is a particular domain park with a bunch of junk pages trying to sell you blenders and stuff" (Interview CC Director). Common Crawl uses some heuristics to identify spam, but in addition to that, spam is the only case where Common Crawl manually monitors and intervenes in the data collection process to prevent the crawler from getting trapped (Interview CC Crawl Engineer).

Staffers at Common Crawl also articulate positive notions about what kinds of data they want to crawl (rather than what they want to avoid). First is language and regional diversity:

> *An important point for me is that the crawl is diverse in terms of language and regional coverage. If I had more time and resources, that would be the first thing I would address.* (Interview CC Crawl Engineer)

Notably, AI builders have requested more language diversity as well (Interview CC Director).

Second is relevance. We allude to this in our description of the crawling process above, which is designed to automatically find new URLs considered good quality. As the main crawl engineer put it, “I need to know that this domain or website is good, the other one less so” (Interview CC Crawl Engineer). Using harmonic centrality scoring, Common Crawl considers a URL to be of higher quality if it is part of a domain that is often linked to (Interview CC Crawl Engineer).

How AI builders filter Common Crawl

Common Crawl's data is meant to be a basis for others to create their own datasets through filtering and categorization. AI builders have created filtered versions for pre-training their LLMs that vary greatly in size, content, and purpose. On a high level, there are four typical filtering approaches (see also Penedo et al. 2023):

- Language filtering: Most popular filtered versions of Common Crawl are English only. AI builders rely on tools to automatically detect languages.
- Keywords and simple heuristics: This can be applied to the URLs, to entire documents, or to lines within documents. For example a line-based rule to only keep lines within a page that end with punctuation marks, or a document-based rule to remove the entire page if it contains certain keywords considered harmful, or similarly remove an entire page if its URL contains harmful keywords.
- AI classifiers: AI builders use another dataset considered high quality (like Wikipedia snapshots) and create a classifier that only keeps URLs from crawls that are statistically similar to this reference (typically using text classifiers with adjustable ranges of similarity thresholds to determine when a page is kept).
- Deduplication: AI builders can remove exact copies of entire documents or parts of documents, or similar to filtering with AI classifiers, they can remove text deemed too similar to other text with adjustable ranges of similarity thresholds.

To illustrate, let's take a closer look at Pile-CC, the version of Common Crawl created by EleutherAI for its LLM training dataset, The Pile (Gao et al. 2020). First, EleutherAI conducted its own text extraction from Common Crawl's WARC files because the provided WET files were deemed too low quality as they included the names of menu items, for example. Second, all non-English content was removed. Third, EleutherAI created an AI classifier and used another dataset it created for The Pile as its high-quality reference: OpenWebText2, which consists of extracted text from all URLs posted on Reddit until April 2020 and upvoted at least three times. The text similarity threshold determining how similar URLs in Common Crawl need to be to OpenWebText2 pages to be included was adjusted to "filter our subset of CC to the size we needed" (Gao et al. 2020).

Filtering techniques in popular versions of Common Crawl have been criticized as inadequate to reliably detect problematic content. EleutherAI itself is a nonprofit AI research lab for open source AI that does not create end user products and is doing pioneering work to make generative AI more transparent and interpretable. However, its version of Common Crawl has been used by many other AI builders without

engaging with the problematic aspects of the filtering methods used to create it. In addition, those filtering methods are not unique to Pile-CC. For instance, using URLs upvoted on platforms with user generated content like Reddit to train reference classifiers is problematic because these users are far from being representative of any population. On Reddit, most users are male and white among other issues of representation bias (Bender et al. 2021). Reddit also struggled with moderating toxic communities for years (Massanari 2017). A portion of URLs shared in those communities has likely been included in OpenWebText2 and similar datasets. In addition to concerns about what constitutes a high-quality reference dataset, excessive filtering with AI classifiers might also decrease the performance of the resulting model. With AI classifier-based filtering, the remaining documents could become too "biased towards the ones with features superficially resembling the high quality data in a way that satisfies the classifier, rather than truly high quality data" (Gao 2021).

The use of simple heuristics and keywords for filtering, which other popular CC versions like Alphabet's C4 solely rely on, has also been criticized. Not only does it fail at fully removing toxic and biased content, but popular keyword lists used to filter AI training datasets like the "List of Dirty, Naughty, Obscene, and Otherwise Bad Words" (used in C4) risk underrepresenting vulnerable groups because those keywords tend to also remove non-toxic content from LGBTQIA+ communities, for example (Dodge et al. 2021; Simonite 2021). At the same time, a lot of data of questionable quality remains, for example thousands of machine-translated Japanese patents in C4 (Dodge et al. 2021).

At the heart of the debate about the adequacy or inadequacy of filtering techniques used in popular Common Crawl versions is an unresolved conflict. Many AI builders insist that LLMs need to be trained on amounts of text data too large for careful manual curation, but given that automated filtering for toxicity and bias has significant limitations (Prabhu and Birhane 2020), they have not established widely adopted guidelines or practices for how much and what kind of problematic content is tolerable (if any) in pre-training data, or how its negative effects on the trained model should be minimized.

Implications of Common Crawl's popularity for trustworthy AI

Common Crawl's newfound popularity among AI builders has had a huge impact on Common Crawl itself. This was most obvious with the redesign of its homepage at the end of August 2023. While the old version did not mention LLMs at all, they are front and center now. For a short time after the redesign, the starting page highlighted that Common Crawl provided "82% of raw tokens used to train GPT-3." A new page about the project's impact was added that exclusively talks about and celebrates the impact of "LLMs based on Open Data," suggesting that Common Crawl's impact exclusively or primarily stems from its role in pre-training LLMs (even though it highlights the large number of research papers using its data elsewhere).

Underlying this change is exponential growth of Common Crawl's user base since the publication of OpenAI's paper on GPT-3 (Brown et al. 2020), with LLM users now overshadowing all others (Interview CC Crawl Engineer). The homepage update reflects the excitement staffers expressed in our interviews about the rising popularity of LLMs:

> *I think for the first 15 years of its existence, Common Crawl has kind of been a sleepy project. It's been cited in over 8,000 research papers and it's been a tremendous resource, but it's really in the past year I think that LLMs have taken off and we're all kind of like, "Oh my God," you know, "What have we done here?" [laughing]* (Interview CC Director)

At the time of our interviews, Common Crawl was also internally in a process of change to better adjust to this new role (Interview CC Director). This has already led to an increase in resources. For a long time, Common Crawl only had one employee and was solely financed by its founder and chairman Gil Elbaz. In 2023, more staff was hired and the organization is actively seeking donations to expand its operation (Interview CC Director).

How all of this will change the way Common Crawl operates is not clear at the time of writing, but based on our interviews it seems likely that Common Crawl intends to stay true to its original vision that less curation of its crawl enables more innovation for downstream users. Asked about Common Crawl's role within the AI ecosystem, the director described it as protecting "the point of ingestion of content":

> *If you say that a human is allowed to read a webpage, but a machine isn't, I think that's a disparity that we would challenge. We think it's important to protect the integrity of*

> *the corpus. If something is on the web, we like to have a copy of it that researchers can easily get access to.... I think that ripping pages out of the internet to try and change what an LLM does is not the right approach. You have to do it at the point of use, not at the point of ingestion.* (Interview CC Director)

With Common Crawl's stance on curation unchanged, its continued influence on LLM development is unlikely to change drastically in the near future. Taking the findings of this report together, let's have a closer look at the implications for trustworthy generative AI.

Trustworthy AI and Common Crawl

Mozilla defines trustworthy AI as AI that is "demonstrably worthy of trust, tech that considers accountability, agency, and individual and collective well-being.... [trustworthy] AI-driven products and services are designed with human agency and accountability from the beginning" (Ricks and Surman 2020). Trustworthy AI products should respect privacy, be transparent in meaningful ways, and give users control over their experience.

When considering the implications of Common Crawl's popularity for trustworthy AI, it is first of all important to emphasize its positive role in helping to make LLM research and development more transparent and competitive. Thanks to its openness, filtered Common Crawl versions used in LLM training are inherently more auditable than any proprietary training datasets. Notably, the two most popular Common Crawl versions in our data collection, Pile-CC and C4, arguably increased the transparency of the field overall. LLMs that are very open and transparent about their data curation and other aspects (like Bloom) typically came from actors outside of big tech (researchers or smaller corporate actors) that do not have the resources to collect the data necessary to build LLMs from scratch. In this way, Common Crawl arguably lived up to its mission and acted as infrastructure for building LLMs that helped to uplift more and often smaller players (relative to big tech). With a very small team and limited resources, the organization manages to maintain a massive amount of freely available data that is unique; there is no alternative for LLM builders without the resources to crawl billions of URLs or pay to access a commercial (and usually less transparent) database.

From the perspective of trustworthy AI, however, Common Crawl's popularity among builders also has downsides. While Common Crawl enables AI builders to be more transparent, they do not necessarily make use of that opportunity. They do not always reveal details about the filtering process and which crawls they extracted from Common Crawl's corpus. Some AI builders, including several who describe their LLMs

as “open,” do not disclose any information about their pre-training data and whether Common Crawl was used (for example like Meta’s Llama v2, see Touvron, Martin, et al. 2023). Moreover, the massive size and diversity of Common Crawl’s data can make it difficult to understand exactly what an LLM has been trained on (Luccioni and Viviano 2021). This is reinforced by the (false) assumption among some AI builders that Common Crawl represents the “entire internet” and by extension, somehow, is a proxy for representing “all human knowledge.” In addition, the filtering techniques AI builders use are often too simplistic to seriously address concerns around toxic and biased training data — something Common Crawl does not provide guidance or leadership on.

The almost symbiotic relationship between Common Crawl and a significant proportion of LLM builders has also led to the training of generative AI on massive amounts of copyrighted material, and it is hotly debated whether they are authorized to do so. As mentioned before, this is sparking “data revolts” (Frenkel and Thompson 2023) by content creators, lawsuits challenging fair use exceptions, and platforms increasingly shutting down or limiting unpaid API access to their data. These developments could trend towards making the internet less open and collaborative in general, and content platforms less transparent and accountable to the public.

When thinking about how the status quo could be improved, it is important to reiterate that Common Crawl provides a basis from which AI builders can create their own (filtered) versions of varying quality. Both Common Crawl as the data source and AI builders as downstream users are equally important to understanding the implications of Common Crawl’s popularity as a pre-training dataset, but individually, each can contribute greatly to improving the trustworthiness of AI. Let’s discuss what each of them can do.

Recommendations for using Common Crawl to train AI

When AI builders use Common Crawl uncritically as though its data covers all or a majority of the entire internet, they essentially declare a relatively small subsection of primarily English web pages as being representative of the entire world, even though it includes proportionately little content from other languages and cultures. Moreover, given the imperfections of automated filtering techniques that leave a lot of problematic content in Common Crawl untouched (and unannotated), including in popular filtered versions like C4 (Schaul, Chen, and Tiku 2023), AI builders have to rely on containing this toxicity by adapting LLMs after pre-training, for example, with fine-tuning techniques like RLHF. To what extent this is possible is questionable (Birhane et al. 2023), but even if we assume that toxic pre-training data can be

sufficiently contained with fine-tuning and other techniques, as of today, keeping LLMs "clean" requires the continued efforts of data workers in often precarious working conditions (Williams, Miceli, and Gebru 2022).

To make generative AI more trustworthy when relying on datasets like Common Crawl, AI builders should either put more effort into filtering Common Crawl for particular types of content and consistently provide proper dataset documentation, or they should put an emphasis on those criteria when choosing a filtered version created by others. As discussed above, filtering for particular types of content in popular Common Crawl versions is often limited to removing pornography and relies on simple keyword lists or AI classifiers trained on user generated content that can itself be problematic (see above). Going forward, AI builders should consider filtering more types of problematic content (for example, content that is racist or misogynist) in more nuanced ways that do not remove non-toxic content created by digitally marginalized communities. Alternatively, some have argued that toxic content in the training data can be used to make the model better at detecting and avoiding it. If such approaches create better results than avoiding harmful content altogether, our suggestions here remain the same, only that problematic content should then be annotated rather than filtered out.

Related to better and more transparent filtering or annotating, there should be a greater diversity of filtered Common Crawl versions for builders to draw from to better reflect the various cultural contexts in which LLMs are deployed (as what is considered toxic or problematic varies across cultural contexts). The notion that there can be a "neutral" or default version of a dataset derived from Common Crawl is a sure path to mirroring systemic inequities of the internet downstream.

A noteworthy example of a filtered Common Crawl version where the authors put a stronger emphasis on filtering is the RefinedWeb dataset by the Technology Innovation Institute (Penedo et al. 2023). Similar to pre-training datasets for early LLMs, RefinedWeb solely relies on Common Crawl, but with more extensive filtering than previous variants. No AI classifiers were used, only keywords and heuristics. To filter out types of content, the authors developed a more nuanced approach that relies on filtering URLs. They combined a list of 4.6M URLs curated by the University of Toulouse and primarily designed to regulate internet usage in schools, and a scoring system for URLs based on the presence of certain keywords. The authors claim that their approach avoids some of the pitfalls of using AI classifiers or simple blocklists that can "overly impact minorities" (Penedo et al. 2023). From a trustworthy AI perspective, RefinedWeb is still far from perfect: The authors only filtered for adult content, and they did not discuss the limitations and biases in Common Crawl's

coverage and how this might influence LLMs trained on RefinedWeb, or the biases in the university's URL block list. Whether RefinedWeb really is less problematic than other Common Crawl versions needs to be independently evaluated. Still, more experimentation like RefinedWeb that explores more nuanced ways to filter Common Crawl is desirable from a trustworthy AI perspective.

When it comes to LLMs for end user products, we also need better industry standards (including through regulation) for evaluating filtered Common Crawl versions and the downstream effects of those versions on models trained on them. Some AI builders already use tools that automatically check for profanity or other notions of problematic content and report the output of those evaluations (like EleutherAI for The Pile, see Gao et al. 2020). More nuanced, and culturally contextual tools that give an impression of the kinds of profanity, racism, discrimination etc. found in the datasets would allow for further customization. Providing a descriptive demographic overview of the diversity of the content that shows, for example, the regions from where the content originates, would be useful for evaluating the dataset's implications for trustworthy AI. Any tool designed to automatically detect abstract and culturally contextual concepts is bound to be imperfect, so ideally there should also be evaluations by human moderators under fair, safe conditions (Davis, Williams, and Yang 2021) before training. Related to filtering for content, clarity on how intellectual property rules apply to training LLMs is important, though this has yet to be solved by regulators worldwide.

In addition, better industry standards and benchmarks for evaluating the effects of individual datasets on model behavior would be very important for assessing the implications of using particular datasets for trustworthy AI. This will require more technical advancements in AI interpretability, but existing tools like EleutherAI's Language Model Evaluation Harness that seek to standardize different evaluation frameworks is a step in this direction.

Finally, another step forward would be trustworthy intermediaries dedicated to filtering Common Crawl for various purposes. At the moment, the authors of popular Common Crawl versions are themselves LLM builders: EleutherAI, Alphabet, the Technology Innovation Institute, and more. This often means that the filtered versions were intended to train particular LLM models, and in all cases filtering techniques are not updated and adjusted to take other use cases or criticisms into account after the original publication. Establishing intermediaries or a commons dedicated to filtering Common Crawl in transparent and trustworthy ways that are continuously improved would help make the LLM ecosystem more transparent and less dependent on datasets that contain a lot of toxicity.

A different Common Crawl?

Even without changing its curation methods, Common Crawl could still influence how AI builders use its data to train their models by changing its Terms of Use or adopting a different data license for its crawl. For example, even just by requiring that the use of Common Crawl's data for training LLMs needs to be attributed, similarly to how Creative Commons Attribution licenses require appropriate credit for republication of photos or text. This could help address the growing lack of transparency when providers of influential LLMs like OpenAI with GPT-4 and Meta AI with Llama v2 refuse to provide any information about their pre-training data. In addition, Common Crawl could require AI builders to document measures taken to ensure their LLMs do not reproduce unauthorized copyrighted materials, or harmful outputs in products.

Common Crawl could also invest in a more curated and values-oriented crawl. At the moment, the discovery and selection of URLs to crawl is almost entirely automated. From a trustworthy AI perspective, this has the disadvantage that only "popular" URLs are included (popular in the sense that they are often linked to), which makes content from digitally marginalized communities less likely to be included (see also Noble 2018 discussing this in relation to Google's PageRank algorithm). This partly is a technical necessity: crawlers need links to find new content. If a link is only posted on a Facebook page or on other social media that Common Crawl is denied access to, it will never discover that URL (Interview CC Crawl Engineer). However, this issue could be mitigated with a community-driven approach, where groups help identify relevant content that would make Common Crawl more diverse, not only in terms of languages and regions, but also in terms of the perspectives represented in its crawls.

Given its mission and history, it seems unlikely that Common Crawl is going to take a more opinionated approach to its crawl. Still, Common Crawl could contribute to making generative AI more trustworthy by acknowledging that its current approach is not neutral, despite what Elbaz envisioned when he described his organizations as "neutral data companies." As Common Crawl does not (and cannot) claim to be representative of the entire web, its samples are necessarily biased. And even if it did crawl the entire web, global digital divides that extend far beyond questions of basic access mean that the web is not representative of all people and measures of popularity or centrality are likely to skew toward expressions of power. Common Crawl could do more to highlight those biases and help users of its data understand the consequences, for example by providing quality and toxicity evaluations, or language labeling. Similarly, being more transparent and inclusive about its governance would

be an important step forward. Common Crawl's lack of transparency here is at odds with its self-image as a public resource. For a long time, there was almost no public communication from the organization outside of its mailing list (which mostly dealt with technical questions) and its blog (mostly dedicated to announcing new crawl data). If Common Crawl was to offer more transparency about data collection and curation decisions, as well as clarity about the project's finances, it would be hugely valuable to AI researchers and help facilitate accountability for what has become a major public resource for technology development. Moreover, establishing community-oriented mechanisms like a formal way to make requests about the crawl or more active participation in community events related to trustworthy AI would help make the project more inclusive. Common Crawl could also provide resources to help AI builders understand the implications of using its data to train LLMs. This can be educational resources about the limitations of its data, or a discussion of the various tools available for filtering and analyzing the data.

Building an alternative to Common Crawl from scratch would be costly, but it would be an opportunity to rethink how to ingest comparably large quantities of data according to a different set of values. Yet another possibility would be to increase the number and diversity of high-quality datasets curated by humans equitably, and sharing them in ways that bring value to the communities that created them (as, for instance, with indigenous data sovereignty). While such datasets may be small relative to Common Crawl individually, a growing number of such datasets might help to reduce the reliance of AI builders on datasets with no human curation and only a bare minimum of automated filtering. Combined with technological advances that make LLMs more performant when trained on less data, this might help create a future where AI builders create LLMs that are interpretable, more transparent, and less harmful — in short, more trustworthy.

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Birhane, Abeba, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. 2023. “On Hate Scaling Laws For Data-Swamps.” arXiv. https://doi.org/10.48550/arXiv.2306.13141.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. “On the Opportunities and Risks of Foundation Models.” *ArXiv*. https://crfm.stanford.edu/assets/report.pdf.

Braun, Virginia, and Victoria Clarke. 2012. “Thematic Analysis.” In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological.*, edited by Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, 57–71. Washington: American Psychological Association. https://doi.org/10.1037/13620-004.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Common Crawl Foundation. 2011. “Answers to Recent Community Questions.” *Common Crawl - Blog* (blog). November 16, 2011. https://commoncrawl.org/blog/answers-to-recent-community-questions.

———. n.d. “Our Mission.” Accessed November 2, 2023. https://commoncrawl.org/mission.

Cremades, Alejandro. 2019. “Gil Elbaz On Google Acquiring His Company And Turning It Into A $15 Billion Business.” Alejandro Cremades. April 4, 2019. https://alejandrocremades.com/gil-elbaz-on-google-acquiring-his-company-and-turning-it-into-a-15-billion-business/.

Davis, Jenny L., Apryl Williams, and Michael W. Yang. 2021. “Algorithmic Reparation.” *Big Data & Society* 8 (2). https://doi.org/10.1177/20539517211044808.

Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. “Bringing the People Back In: Contesting Benchmark Machine Learning Datasets.” arXiv. https://doi.org/10.48550/arXiv.2007.07399.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.” arXiv. https://doi.org/10.48550/arXiv.2104.08758.

Frenkel, Sheera, and Stuart A. Thompson. 2023. “‘Not for Machines to Harvest’: Data Revolts Break Out Against A.I.” *The New York Times*, July 15, 2023, sec. Technology. https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html.

Gao, Leo. 2021. “An Empirical Exploration in Quality Filtering of Text Data.” arXiv. https://doi.org/10.48550/arXiv.2109.00698.
Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, et al. 2020. “The Pile: An 800GB Dataset of Diverse Text for Language Modeling.” arXiv. https://doi.org/10.48550/arXiv.2101.00027.
Hayes, Brian. 2015. “Crawling toward a Wiser Web.” *American Scientist* 103 (3): 184. https://doi.org/10.1511/2015.114.184.
Kim, Sung. 2023. “List of Open Sourced Fine-Tuned Large Language Models (LLM).” *Medium* (blog). September 30, 2023. https://sungkim11.medium.com/list-of-open-sourced-fine-tuned-large-language-models-llm-8d95a2e0dc76.
Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, et al. 2023. “The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset.” arXiv. https://doi.org/10.48550/arXiv.2303.03915.
Leetaru, Kalev. 2017. “Common Crawl And Unlocking Web Archives For Research.” Forbes. September 28, 2017. https://www.forbes.com/sites/kalevleetaru/2017/09/28/common-crawl-and-unlocking-web-archives-for-research/.
Luccioni, Alexandra Sasha, and Joseph D. Viviano. 2021. “What’s in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus.” arXiv. https://doi.org/10.48550/arXiv.2105.02732.
Massanari, Adrienne. 2017. “#Gamergate and The Fappening: How Reddit’s Algorithm, Governance, and Culture Support Toxic Technocultures.” *New Media & Society* 19 (3): 329–46. https://doi.org/10.1177/1461444815608807.
McKinsey. 2023. “What Is Generative AI?” January 19, 2023. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai.
Morrison, Sara. 2023. “The Tricky Truth about How Generative AI Uses Your Data.” Vox. July 27, 2023. https://www.vox.com/technology/2023/7/27/23808499/ai-openai-google-meta-data-privacy-nope.
Munn, Luke. 2022. *Countering the Cloud: Thinking With and Against Data Infrastructures*. 1st ed. New York: Routledge. https://doi.org/10.4324/9781003341185.
Nagel, Sebastian. 2019. “Commoncrawl vs Archive.Org Etc.” *Common Crawl Mailing List*. https://groups.google.com/g/common-crawl/c/RBFAn0o55cY/m/68qiLwZMBAAJ.
———. 2022. “Questions about Using Common Crawl for Another Hugging Face Project.” *Common Crawl Mailing List*. https://groups.google.com/g/common-crawl/c/BgPvP6HB2n0/m/P-Nw5YoJAQAJ.
———. 2023. “Common Crawl: Data Collection and Use Cases for NLP.” Presented at the HPLT & NLPL Winter School on Large-Scale Language Modeling and Neural Machine Translation with Web Data, February 6. http://nlpl.eu/skeikampen23/nagel.230206.pdf.
Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce*

Racism. New York: New York university press.
Orr, Will, and Kate Crawford. 2023. “The Social Construction of Datasets: On the Practices, Processes and Challenges of Dataset Creation for Machine Learning.” https://doi.org/10.31235/osf.io/8c9uh.
Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. “Training Language Models to Follow Instructions with Human Feedback.” arXiv.Org. March 4, 2022. https://arxiv.org/abs/2203.02155v1.
Owens, Trevor. 2014. “Machine Scale Analysis of Digital Collections: An Interview with Lisa Green of Common Crawl – Coffeehouse.” January 29, 2014. https://coffeehouse.dataone.org/2014/01/29/machine-scale-analysis-of-digital-collections-an-interview-with-lisa-green-of-common-crawl/.
Patel, Nilay. 2023. “Microsoft CTO Kevin Scott Thinks Sydney Might Make a Comeback.” The Verge. May 23, 2023. https://www.theverge.com/23733388/microsoft-kevin-scott-open-ai-chat-gpt-bing-github-word-excel-outlook-copilots-sydney.
Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. “The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only.” arXiv. https://doi.org/10.48550/arXiv.2306.01116.
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. http://www.aclweb.org/anthology/D14-1162.
Prabhu, Vinay Uday, and Abeba Birhane. 2020. “Large Image Datasets: A Pyrrhic Win for Computer Vision?” arXiv. https://doi.org/10.48550/arXiv.2006.16923.
Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” arXiv. https://doi.org/10.48550/arXiv.1910.10683.
Rampin, Rémi, and Vicky Rampin. 2021. “Taguette: Open-Source Qualitative Data Analysis.” *Journal of Open Source Software* 6 (68): 3522. https://doi.org/10.21105/joss.03522.
Ricks, Becca, and Mark Surman. 2020. “Creating Trustworthy AI. A Mozilla Whitepaper on Challenges and Opportunities in the AI Era.” Draft V1.0. Mozilla Foundation. https://foundation.mozilla.org/en/insights/trustworthy-ai-whitepaper/.
Rogers, Bruce. 2014. “Gil Elbaz Builds Factual To Be The World’s Data Steward.” Forbes. May 29, 2014. https://www.forbes.com/sites/brucerogers/2014/05/29/gil-elbaz-builds-factual-to-be-the-worlds-data-steward/.
Schaul, Kevin, Szu Yu Chen, and Nitasha Tiku. 2023. “Inside the Secret List of Websites That Make AI like ChatGPT Sound Smart.” Washington Post. April 19, 2023. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.
Simonite, Tom. 2021. “AI and the List of Dirty, Naughty, Obscene, and Otherwise Bad Words.” *Wired*, February 4, 2021.

https://www.wired.com/story/ai-list-dirty-naughty-obscene-bad-words/.
Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. “LLaMA: Open and Efficient Foundation Language Models.” arXiv. https://doi.org/10.48550/arXiv.2302.13971.
Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” arXiv. https://doi.org/10.48550/arXiv.2307.09288.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” arXiv. https://doi.org/10.48550/arXiv.1706.03762.
Williams, Adrienne, Milagros Miceli, and Timnit Gebru. 2022. “The Exploited Labor Behind Artificial Intelligence.” *Noema*, October 13, 2022. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence.
Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.” arXiv. https://doi.org/10.48550/arXiv.2304.13712.
Zaino, Jennifer. 2012. “Common Crawl Founder Gil Elbaz Speaks About New Relationship With Amazon, Semantic Web Projects Using Its Corpus, And Why Open Web Crawls Matter To Developing Big Data Expertise.” *DATAVERSITY* (blog). January 20, 2012. https://dev.dataversity.net/common-crawl-founder-gil-elbaz-speaks-about-new-relationship-with-amazon-semantic-web-projects-using-its-corpus-and-why-open-web-crawls-matter-to-developing-big-data-expertise/.
Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. “A Survey of Large Language Models.” arXiv. https://doi.org/10.48550/arXiv.2303.18223.