A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl

Stefan Baack
Mozilla Foundation
stefan@mozillafoundation.org

ABSTRACT

Common Crawl is the largest freely available collection of web crawl data and one of the most important sources of pre-training data for large language models (LLMs). It is used so frequently and makes up such large proportions of the overall pre-training data in many cases that it arguably has become a foundational building block for LLM development, and subsequently generative AI products built on top of LLMs. Despite its pivotal role, Common Crawl itself is not widely understood, nor is there much reflection evident among LLM builders about the implications of using Common Crawl's data. This paper discusses what Common Crawl's popularity for LLM development means for fairness, accountability, and transparency in generative AI by highlighting the organization's values and practices, as well as how it views its own role within the AI ecosystem. Our qualitative analysis is based on in-depth interviews with Common Crawl staffers and relevant online documents.

After discussing Common Crawl's role in generative AI and how LLM builders have typically used its data for pre-training LLMs, we review Common Crawl's self-defined values and priorities and highlight the limitations and biases of its crawling process. We find that Common Crawl's popularity has contributed to making generative AI more transparent to scrutiny in many ways, and that it has enabled more LLM research and development to take place beyond well-resourced leading AI companies. At the same time, many LLM builders have used Common Crawl as a source for training data in ways that are problematic: for instance, with lack of care and transparency for how Common Crawl's massive crawl data was filtered for harmful content before the pre-training, often by relying on rudimentary automated filtering techniques. We offer recommendations for Common Crawl and LLM builders on how to improve fairness, accountability, and transparency in LLM research and development.

CCS CONCEPTS

• **Information systems** → World Wide Web; Web searching and information discovery; Web search engines; Web crawling; • **Computing methodologies** → Artificial intelligence.

KEYWORDS

Training data, Pre-training, Large language models, Generative AI

ACM Reference Format:
Stefan Baack. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3630106.3659033

1 INTRODUCTION

Common Crawl (CC) has become an important enabler of generative AI that is virtually unknown to the broader public. It is a massive archive of web crawl data amassed by a handful of people working for a small California nonprofit since 2008. Thanks to its size, diversity, and free of charge availability on Amazon Web Services, it has become a very popular source of training data for LLM builders. GPT-3, the LLM powering the first version of ChatGPT published in 2022, would not have been possible without Common Crawl, as more than 80% of its tokens came from it [7]. A majority of LLMs published both before and since then by other developers likewise rely heavily on filtered versions of Common Crawl for their pre-training.

Despite its popularity among LLM builders, Common Crawl's role in generative AI has received relatively little attention or scrutiny. When it does appear in discussions, it is often framed as a "copy of the internet," or of (nearly) the "entire internet" because of its size (see for example [21, 22]). Such claims further obscure Common Crawl's prevalent role in generative AI because they ignore the limitations and biases of its data. Common Crawl was founded in 2007 and its mission was never centered on providing AI training data but rather on leveling the playing field for tech development outside of the biggest internet companies. LLM builders only have become its main user group in recent years, but for most of its history the majority of users were researchers from various fields. And while the size of its archive with more than 9.5 petabytes is enormous, Common Crawl emphasizes that it is not the "entire web," nor even a representative sample of it.

How LLM builders use Common Crawl is crucial for the fairness, accountability, and transparency of their models. Training data has direct implications for model behavior and who is likely to be empowered and disempowered by applications built on top of those models. In this paper, we contribute to a more reflective and critical use of Common Crawl by examining its values and practices in-depth to explore the implications of its popularity among LLM builders. Based on this analysis, we consider how this popularity has been both beneficial and detrimental to fair, accountable, and transparent LLM research and development. In more detail, this paper makes the following contributions:

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659033

- In chapter 2 and 3, we present our approach to studying datasets like Common Crawl as socio-technical infrastructure and discuss how this adds to the existing critical research on AI training datasets and their downstream effects on models.
- In chapter 4, we discuss how LLM builders typically use Common Crawl and highlight that the popularity of Common Crawl has shaped builders' expectations regarding model behavior.
- In chapter 5 and 6, we provide an analysis of Common Crawl's mission and its crawling process, highlighting the biases and limitations of the data it provides while pushing back on the false narrative of LLMs being trained on (a representative sample of) the "entire internet."
- In chapter 7, we highlight that Common Crawl's mission and purpose as an organization does not neatly align with the needs of fair and accountable LLM development, and outline steps Common Crawl and LLM builders could take to make generative AI more fair, accountable, and transparent.

2 RESEARCHING AI DATASETS

Researchers have frequently demonstrated representation bias in AI training data which is connected to discriminatory outputs by AI models trained on it (see for example [40]). Some have also investigated Common Crawl's corpus. For example, [19] found that Common Crawl contains a "significant amount of undesirable content, including hate speech and sexually explicit content." Similarly, [5] analyzed the contents of LAION-400M, a popular training dataset consisting of Image-Alt-text pairs parsed from Common Crawl. They showed that a significant amount of offensive content remains in filtered Common Crawl versions including those shared by the non-profit Large-scale Artificial Intelligence Open Network (LAION).

Instead of auditing AI training datasets directly, some researchers have instead examined them as socio-technical constructions to help explain "why datasets embody specific political perspectives or offer insights into how the practices of dataset creation can be improved" [28]. For example, [3] demonstrated that highly cited machine learning papers overwhelmingly emphasized values related to technical performance, while societal needs and negative potential are rarely addressed in comparison. [39] similarly analyzed values expressed in dataset documentation in computer vision, highlighting both what values were made explicit, as well as what contrasting values were "silenced." They found that "computer vision dataset practices value *efficiency over care*, *universality over contextuality*, *impartiality over positionality*, and *model work over data work*." [29] took a different approach and studied a popular AI training dataset, a filtered Common Crawl snapshot called "C4" (see below), and concluded that dataset construction is an under-researched practice mostly based on ad hoc processes shaped by "pressures of *scale*, the struggle for *resources*, the adoption of *shortcuts*, and confusion about *accountability*."

In this paper, we complement [29] by studying Common Crawl itself as *infrastructure* for LLM development. Unlike the specific datasets examined by the authors cited above, LLM builders do not train their models directly on Common Crawl's archive because it contains too much content deemed undesirable (see chapter 4). LLM builders have to filter Common Crawl before the training — or, more often, they use or recreate filtered versions created by others, like EleutherAI's Pile-CC [14]. Common Crawl itself is therefore *not* an LLM training dataset, but it has an infrastructural role in LLM research and development as a foundation for creating training data. Infrastructures are not neutral mediators, they privilege certain ways of thinking and doing things:

> "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." [23]

Following this framing of Common Crawl as infrastructure, our goal was to analyze the organization's stated values and practices, and what they imply for LLMs trained on its data.

3 METHODS AND LIMITATIONS

To study the influence of Common Crawl as infrastructure for LLM research and development, we adapted [11]'s genealogical approach of investigating "how and why [machine learning] datasets have been created, what and whose values influence the choices of data to collect, the contextual and contingent conditions of their creation." We primarily relied on the qualitative analysis of interviews and online materials to examine the motivations, assumptions, and values of staffers in leading positions at Common Crawl.

We conducted semi-structured interviews with Common Crawl's director and main crawl engineer in mid-2023. Questions to both interviewees related to how Common Crawl reflects the web, the crawling process, data curation, and Common Crawl's relationship with LLM builders. The interview with the director focused more on the organization's mission and purpose, while the crawl engineer was asked more about the crawling process and of how it changed over time. Each interview was about 40 minutes long.

Because Common Crawl had a small team at the time of our study, we additionally collected the following documents:

- Eight discussion threads from Common Crawl's public mailing list at groups.google.com/g/common-crawl: After examining 376 discussion threads from January 2020 (the year OpenAI published GPT-3) to October 2023, we selected individual posts by staffers about Common Crawl's coverage of the web, the relationship with LLM builders, its mission, and technical explanations of the crawling process. Informed by those posts and our interviews, we searched the archives going back to 2011 with 1157 discussion threads with the following keywords: [LLM], [LLMs], ["large language model"], [OpenAI], ["Open AI"], [C4], ["Colossal Clean Crawled Corpus"], [AI], [NLP], [quality], [blekko][1].
- Common Crawl's website at commoncrawl.org: All web pages and three relevant posts from the blog (which dates back to 2011) were included.
- Presentation slides created by the main crawl engineer: The presentation provided details on the crawling process and the

[1] blekko was a search engine startup that played an important role in Common Crawl's history as one of the biggest seed donors. See chapter 6.1.

organization's history for builders in the natural language processing (NLP) field. See [26].

- Three published interviews with Common Crawl's founder and chairman Gil Elbaz: To better understand 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 included [9, 37, 47] in our analysis.
- Statements by past leadership in two articles: Lisa Green [31] and Sara Crouse [18], two former directors of Common Crawl commented on its historical trajectory and self-perception.

We analyzed this data using inductive qualitative coding following thematic analysis [6]. We did not use any text classification models and conducted our analysis with the open source qualitative research software Taguette [36].

Our analysis has some deliberate limitations. First and most importantly, we did not interview LLM builders about their use of Common Crawl. Our primary interest was in Common Crawl's self-described values and practices and thus we limited ourselves to look at typical filtering techniques described in publications by LLM builders. We see our work as complementary to [29] and consider it a useful basis for future research into further evaluating Common Crawl's role in generative AI. Second, we did not analyze Common Crawl's early history in depth. Most LLM builders train their models on snapshots of Common Crawl that were crawled 2017 or later, so our focus was on understanding how the crawls were conducted within this timeframe. LLMs trained on significant amounts of pre-2017 crawls might be affected by the longer evolution of Common Crawl's approach to crawling the web. Given Common Crawl's continued relevance for LLM research and development, a more comprehensive analysis of its full archive would be another valuable follow-up project.

4 COMMON CRAWL'S ROLE IN GENERATIVE AI

Common Crawl has long been popular in NLP [16, 33], but the invention of the transformer technology [42] that powers modern generative AI created new demand for large quantities of diverse training data. This made Common Crawl more relevant, as its archive is massive and made up almost entirely of text spread across billions of URLs. However, a challenge for LLM builders is the large amount of content in Common Crawl undesirable for model training: hate speech, explicit and abusive content, and other types of problematic material [19], as well as "boiler-plate text like menus, error messages, or duplicate text" [35]. Therefore, LLM builders train their models on filtered samples of Common Crawl's archive and there are a handful of filtered versions that are reused frequently, especially Alphabet's "Colossal Clean Crawled Corpus" ("C4," see [35]) and EleutherAI's "Pile-CC" (which is part of EleutherAI's LLM training dataset "The Pile," see [14]). Typical filtering techniques include (see also [32]):

- Language filtering: LLM builders typically create language specific subsets of Common Crawl. The most popular versions are English only.
- Keywords and simple heuristics: It is common to remove pages that contain keywords considered harmful in the URL or anywhere within the page. An example for a heuristic is a line-based rule to only keep lines ending with a punctuation mark.
- AI classifiers: A reference dataset considered high quality (for instance Wikipedia text data) is used to train a text classifier. This classifier is used to filter out everything from Common Crawl that does not meet an adjustable similarity threshold.
- Deduplication: This can involve removing completely identical pages or sections of text within pages. Classifiers can also be used here to remove text that is too similar.

As an example, Pile-CC used a mix of these techniques [14]. The authors removed non-English content and created an AI classifier with OpenWebText2 as its high-quality reference. OpenWebText2 is also part of The Pile and contains the text of all URLs shared and upvoted at least three times on Reddit until April 2020. This classifier was used to "filter our subset of CC to the size we needed" [14] by setting the similarity threshold for determining what pages in Common Crawl to keep accordingly.

Since 2020, most LLM builders compose their training data with two types of datasets: targeted sourced data and broad chunks of web crawl data. Targeted sources are relatively small individually, relatively well defined, and chosen because they are considered "high quality" and diverse to help the LLM imitate various styles of language. Examples are Wikipedia snapshots, arXiv for scientific text, Project Gutenberg for books and more. Using multiple targeted sources was found to potentially "improve the general cross-domain knowledge and downstream generalization capabilities of the model" [14]. However, most LLM builders do not solely rely on targeted sources because the overall size and diversity of the training data is not deemed sufficient to reach the desired performance of the models. Therefore, web crawl data continues to make up significant amounts of the training data in most cases. Some leading AI companies, like Alphabet, Microsoft, Meta, and more recently also OpenAI, have their own crawlers to collect this data themselves. Almost everyone else, at least those that publicly share information about their LLM training data, appears to be using Common Crawl, which continues to make up significant proportions of the training data for many models, for example in GPT-3 [7], LLaMA v1 [41], Falcon [32], or Pythia [2].

Large chunks of Common Crawl are used so frequently that LLM builders include them to ensure model performance is comparable. An example is BigScience, a one-year collaborative research workshop from 2021-2022 aimed at creating more open multilingual LLMs and datasets compared to those of leading AI companies. Part of these efforts was the curation of a training dataset called the "ROOTS Corpus," which contained a filtered Common Crawl version because "not including it would constitute too much of a departure and risk invalidating comparisons [with previously released LLMs]" [17].

However, the way LLM builders filter Common Crawl, especially in the more popular filtered versions, has been criticized for failing to remove problematic content as well as hurting the representational diversity of the training data. With AI classifiers, the issue is how "high quality" is defined. For instance, Reddit upvotes are problematic quality measures since Reddit users are largely homogenous and include many participants of toxic communities [1, 20]. At least some of the URLs upvoted by such toxic communities are likely included OpenWebText2 or similar datasets. Moreover, applying AI classifier filtering too aggressively creates datasets with documents "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" [13]. The way other filtered versions like C4 relied on heuristics and keywords is problematic as well. Among other things, C4 removed pages containing words in a popular crowdsourced keyword list called the "List of Dirty, Naughty, Obscene, and Otherwise Bad Words."[2] This list is focused on eliminating pornography and has also been shown to remove innocuous content from LGBTQIA+ communities [12].

Given the inherent limitations of automatically detecting and removing problematic content [34], there is a need for better industry standards and best practices for filtering sources like Common Crawl that contain such content. If LLM builders insist on using such sources, there needs to be a discussion on how much harmful content is tolerable, who is responsible, and how to deal with the consequences of training models on it. As we discuss below, Common Crawl's popularity makes such guidelines and best practices even more urgent because its mission and subsequently its data do not easily align with fair and accountable LLM development.

5 COMMON CRAWL'S MISSION: BECOMING A PUBLIC RESOURCE FOR OPEN-ENDED RESEARCH AND INNOVATION

Common Crawl's mission and values have been shaped by one individual: Gil Elbaz. He founded the organization in 2007 and has also financed it and acted as its chairman ever since. This makes him the only person that has been involved continuously throughout Common Crawl's history.

As Elbaz has said in the past, founding Common Crawl was motivated by his experience at Google (now Alphabet). Initially, he was one of the co-founders of Applied Semantics, a company which developed the software AdSense for serving contextual advertisements on websites. Google acquired Applied Semantics in 2003, and Elbaz worked for Google until 2007. Later, he explained that the reason for his departure was his concern that Google was becoming a "monopoly of innovation" thanks to the huge amounts of data it has access to:

> "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 [9])

The "neutral data company" Elbaz referred to creating above was Factual, a for-profit company he founded in 2008. Factual sold access to cleaned and structured location data to companies like Microsoft before merging with Foursquare in 2020.[3] Factual's purpose as the second organization Elbaz founded after leaving Google is instructive to further understand Common Crawl's role. Given Common Crawl's mission to give "small startups or even individuals" access to "high quality crawl data that was previously only available to large search engine corporations," [8] it can be considered another attempt at a "neutral data company." While Factual's services as a for-profit business was cleaning and structuring web data, Common Crawl as a nonprofit provides largely unrefined web crawl data for free, but requires downstream users to further curate, annotate, and structure it. This makes Common Crawl act more like an infrastructure than a data service. Common Crawl's director similarly emphasized that the purpose of the project is to jump-start initiatives:

> "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 absence of content curation or moderation is framed as vital to this infrastructural quality. As the director put it, less curation enables more research and open innovation by downstream users:

> "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)

Because Common Crawl avoids curating and annotating its data, its influence on LLM research and development is mostly determined by how it collects its data, i.e. how it decides what to crawl and with what limitations.

6 COMMON CRAWL'S DATA: MACHINE SCALE ANALYSIS

Common Crawl's target audience are "data users," i.e. "programmers, data scientists, researchers working with web data" [24]. Its goal is to enable the automated analysis of web data spread across domains rather than in-depth investigations of individual domains. Referring to this goal, former director Lisa Green described Common Crawl as an enabler of "machine scale analysis" (in [31]).

Common Crawl's data collection is a compromise between enabling "machine scale analysis" while staying within the US fair

[2] See https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.

[3] See https://www.wsj.com/articles/foursquare-merges-with-factual-another-location-data-provider-11586193000.

use doctrine [18]. This means it only collects HTML code (images, PDFs and other media are rarely included)[4], and it provides its data in ways that makes consumption of individual pages difficult (WARC files containing HTML code or WET files with extracted plain text that have to be accessed via AWS). Moreover, it only collects a sample of pages from each domain it crawls. Meaning there will only be some Wikipedia articles in Common Crawl, but not a full copy of the entire website. Which domains to crawl and how many pages of those domains to include depends on their "harmonic centrality" (see next section). Another important aspect is that Common Crawl's archive consists of many datasets containing crawl data, rather than just one huge database. LLM builders have to select which crawls to include and then download and combine them to create training datasets.

Common Crawl also offers a separate "News Crawl" that is updated more frequently. Because this data appears to be far less frequently used by LLM builders,[5] we will concentrate on the main crawl in this paper.

6.1 How Common Crawl decides which URLs to crawl

Common Crawl's approach to crawling the web went through several iterations since 2007 [26]. A challenge for the project has been to continuously discover suitable new URLs to crawl from the vastness of the web. For some time, the project mostly relied on external "seed donations," i.e. annotated lists of URLs shared by other companies. As these contributions dwindled over time, Common Crawl has almost entirely automated its crawls by relying on "harmonic centrality" since 2017 (Interview CC crawl engineer).

Similar to Google's PageRank algorithm, harmonic centrality is a mathematical method for measuring the importance of nodes in a network. In PageRank, the importance of a node depends on the importance of other nodes linked to it (this is referred to as eigenvector centrality). Harmonic centrality, however, is "flat" in that it treats all nodes as equal. It is a variant of closeness centrality where the "closer" a node is to all other nodes, the more central it is. Applied to web pages, it means that the importance of a domain is determined by how many direct and indirect links exist to it from other domains. Direct links are considered "closest" and contribute the most to a domain's centrality score. The more "distant," i.e. indirect a link connection is, the less it contributes. The higher the overall harmonic centrality score of a domain, the more likely it is to be included in a crawl, and the greater number of its pages are fetched. In other words, domains with high scores are more likely to be crawled and have more of their pages included in Common Crawl compared to lower scoring domains. Common Crawl only calculates a harmonic centrality score per domain. To guide the selection of pages from domains, the domain score is projected to all its pages in order to rank them. According to Common Crawl, harmonic centrality is better for avoiding spam than other centrality measures (Interview CC crawl engineer).

Internally, Common Crawl relies on a database called CrawlDB, which records, among other things, the harmonic centrality score and a timestamp of the last attempted crawl for each URL. This database contained 25 billion URLs in August 2023 (Interview CC crawl engineer) and is expanded each month with URLs from new main crawls, smaller crawls dedicated to finding new URLs, and an analysis of sitemaps [25]. To sample URLs from CrawlDB, the harmonic centrality score is modified to give higher priority to new URLs, or URLs that have not been crawled in a while by increasing or reducing scores depending on when the URL was last crawled successfully. There is not a fixed minimum score for inclusion, the threshold is flexible and depends on a monthly quota (Interview CC crawl engineer).

6.2 Limitations and quality measures

Common Crawl's mission is to create "high quality crawl data" [8]. The monthly limit for how many URLs to crawl has thus been set to balance size and quality:

> "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)

The emphasis on quality might seem to contradict Common Crawl's stance on not curating the contents of the data it collects. However, Common Crawl relies on quantifiable and thus measurable notions of quality that are aligned with this stance: language and regional coverage and diversity, and a statistical definition of relevance (harmonic centrality, as described above).

Throughout our interviews, staffers emphasized that they want the crawl to be diverse "in terms of language and regional coverage" (Interview CC crawl engineer). However, a challenge for Common Crawl's coverage of the web is not knowing the size of the web as a whole. The director claims 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 main crawl engineer echoed this uncertainty about the size of the web: "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). While staffers are confident that their crawl is "reasonably representative" (Interview CC crawl engineer), they are uncertain how their archive is representative of the web and reject the idea that Common Crawl is a "copy of the internet":

> "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)

There are a couple of limitations for how much of the web Common Crawl is able to cover:

- Around 50% of the URLs crawled monthly have been included previously (Interview CC crawl engineer). In part,

[4] See Common Crawl's statistics at https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes.

[5] See HuggingFace statistics at https://huggingface.co/models?dataset=dataset:cc_news.

this is due to the reliance on harmonic centrality, as some top domains like Wikipedia always score high, while lower scoring domains are less likely to be included.
- The monthly crawls are discontinuous, and do not take into account content that was published in between.
- In order to restrict AI companies from using their content as training data, a growing number of rights holders, including big and important domains like the New York Times, now block Common Crawl from visiting most if not all their pages using a technical standard for blocking web crawlers called robot.txt, which Common Crawl respects. Big social media platforms like Facebook, that make up significant portions of the web, have been blocking Common Crawl long before this controversy.
- The majority of content in Common Crawl is English, in part because all the technical infrastructure is based in the United States, which creates a bias for English, for example because multilingual pages tend to serve the English version by default.

Instead of a "copy of the internet," the director describes Common Crawl as an "academic sampling of the web" (Interview CC director). It is a sample in the sense that Common Crawl as a whole is a selection of web pages that does not claim to be representative, and individually each monthly crawl is a sample of URLs taken from the internal CrawlDB.

A noteworthy exception to Common Crawl's stance on not curating content is spam, which is considered undesirable and can also damage the diversity of the crawl. Domain parks can keep crawlers trapped in a net of interconnected spam pages. If undetected, the majority of a crawl would end up coming from a domain park consisting of "junk pages trying to sell you blenders and stuff" (Interview CC director). Because it is difficult to reliably detect domain parks automatically, Common Crawl monitors and manually intervenes in the crawling process to make sure the crawler is not getting stuck (Interview CC crawl engineer).

7 IMPLICATIONS OF COMMON CRAWL'S POPULARITY FOR THE FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY OF LLM RESEARCH AND DEVELOPMENT

Since the publication of OpenAI's paper on GPT-3 [7], Common Crawl experienced an exponential growth of its user base thanks to the influx of LLM builders (Interview CC crawl engineer). Staffers are excited about this newfound relevance:

> "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)

The growth in user base also resulted in a growth in resources and staff for Common Crawl. Around the time GPT-3 was published, Common Crawl had only one employee, and the organization was solely financed by its founder and chairman Gil Elbaz. Since our interviews, Common Crawl has raised funds and hired more staff. Despite these changes, our interviews suggest Common Crawl's stance on minimal curation in the name of enabling more research and innovation is not about to change. Its director describes Common Crawl's role in the AI ecosystem 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)

Below, we consider the implication of this for the fairness, accountability, and transparency of LLM research and development.

7.1 Common Crawl and LLM builders: Shared responsibility for fairness, accountability, and transparency

Common Crawl's commitment to openness and transparency has played a positive role in helping to make LLM research and development more transparent and competitive. The popularity of filtered versions like Pile-CC and C4 (that provide a lot of details about their filtering techniques) over less transparent and audible proprietary datasets enabled more scrutiny and more LLM research and development outside a handful of resourceful tech companies. In fact, the most open and transparent LLMs (like Bloom [46]) were built by researchers or smaller corporate actors without the resources to collect terabytes of web crawl data. This is arguably in line with Common Crawl's mission to uplift more builders to compete in this space.

The negative implications of Commons Crawl's popularity partly stem from how LLM builders have used the data, but also from the fact that Common Crawl's mission does not easily align with fair and accountable LLM development. Common Crawl *wants* its data to contain problematic content to enable open-ended research and innovation, but it does not want to take responsibility for annotating it. For LLM development, this is a problematic starting point that requires careful consideration for how Common Crawl's data is filtered before any model training. While some of the most popular filtered versions are transparent, not all LLM builders likewise share which crawls were taken from Common Crawl's archive or provide enough information on how they filtered the data. Leading AI companies increasingly do not disclose any information about the pre-training data for "competitive reasons" (like OpenAI with GPT-4 [43] or Meta AI with LLama v2 [48]). A possible justification for this lack of transparency and documentation could rest on false (and potentially self-serving) assumptions that Common Crawl captures the "entire web," or at least a reasonably representative sample of it. In cases with enough transparency to audit the data, the other big problem is that common filtering techniques have severe shortcomings, which numerous studies have demonstrated (see above).

Common Crawl and LLM builders have a shared responsibility to make LLM development more fair, accountable, and transparent. In the following, we discuss how they can each help to improve the status quo.

7.1.1 Possible contributions by Common Crawl. Highlighting limitations and biases. Because it cannot crawl the entire web, nor be certain about the representativeness of its data, Common Crawl's samples are necessarily biased. Common Crawl could document those biases more prominently as well as the risks for LLMs trained on its data. Among other things, this can include clearly indicating the limitations of the crawl or sharing the results of various quality and toxicity evaluations (and perhaps supporting the development of auditing tools and research). Common Crawl could also provide educational resources tailored to LLM builders for creating training datasets based on its archive in responsible ways.

More transparent and inclusive governance. Common Crawl's nonprofit mission positions the organization as a public resource, but until recently, there was very little public communication by the project. Especially decisions about its data collection and curation should be prominently communicated and justified. Going further, the governance structure should be more inclusive, for example by establishing formal ways for requesting changes to the crawl. Finally, becoming a more active participant in events and discussions related to ethical AI development would help the project be more inclusive and responsive to a wider variety of stakeholders.

Changing its terms of use to enforce more transparency. Common Crawl could change its Terms of Use or adopt a new data license to require attribution (similar to the Creative Commons Attribution licenses) or even disclosure of filtering mechanisms when its data is used to train LLMs. This would help to counter a lack of transparency especially among leading AI companies to not disclose information about the training data (see above).

Less automation and a more curated approach. As a more expansive deviation from its current mission and values, Common Crawl could take steps to have its crawl be more curated, for example with a crowdsourcing approach. The way Common Crawl has automated its operation for URL discovery makes content from digitally marginalized communities less likely to be included, which is problematic for fairness in generative AI (cf. [27] about similar issues with Google's PageRank). A more collaborative approach could help to mitigate these biases by improving language and regional coverage, as well as covering more perspectives of what on the web "matters," i.e. is included.

7.1.2 Additional steps for LLM builders. The way LLM builders typically use Common Crawl leads to two major problems. First, false assumptions about Common Crawl being a "copy of the internet" take a relatively small selection of primarily English web pages as representative for the entire global population, obscuring problems of representativeness and fairness. Second, common pre-training data curation practices fail to remove or annotate significant amounts of problematic content in Common Crawl (cf. [19, 38]), which leads to a greater reliance on containing toxicity with fine-tuning techniques like reinforcement learning from human feedback [30]. It is unclear if sufficient containment is even possible given the growing size of LLMs [4]. Moreover, for leading AI companies offering general purpose applications, this approach means keeping LLMs safe requires moderation by data workers, often under precarious working conditions [45]. Here, we just highlight some high level steps to start addressing those issues.

More effort for filtering problematic content. The most popular filtered versions today primarily remove pornographic content or use AI classifiers based on user generated content, both of which can further harm the representation and experiences of digitally marginalized communities. There is a need for more nuanced filtering efforts that go beyond pornographic content. These filtering techniques should consistently be documented to help LLM builders choose a filtered version for their models. The obvious difficulty of determining what is problematic depending on context and culture does not justify minimal effort.

Create or support dedicated filtering intermediaries. Filtering Common Crawl responsibly is difficult and time consuming. Especially when it comes to removing various types of harmful content, it requires constant effort. What is considered harmful and offensive in one place and time can have a completely different meaning in another context. Currently, the most popular filtered Common Crawl versions were created by LLM builders themselves (like EleutherAI) as a step towards their actual goal: training LLMs. This restricts the amount of time and energy that can be dedicated to the filtering effort, and it means that the filtering techniques are not updated after the publication to take criticism and feedback into account. A possible way forward would be to foster an ecosystem of dedicated filtering intermediaries that can be tasked with continuously filtering Common Crawl in transparent and accountable ways. Establishing and supporting such intermediaries would make the LLM ecosystem less dependent on a handful of filtered Common Crawl versions that contain significant amounts of problematic content.

Less reliance on a handful of Common Crawl versions. Rather than defaulting to Pile-CC, C4 or other versions, there should be a greater variety of filtered Common Crawl versions builders use. Even with better filtering, relying on only a handful of "default" versions risks amplifying global inequalities in LLM research and development. There already are notable attempts to filter Common Crawl in various ways, like [32]'s extensive filtering aimed at removing harmful content without negatively impacting minority groups, or [44]'s work on creating monolingual datasets. Such efforts should be expanded.

Establish industry standards and best practices to reduce harms when using Common Crawl or similar sources. Especially for LLMs used in end-user products, better industry standards (and potentially regulation) for assessing filtered Common Crawl versions are needed. This can include investment into new, and culturally contextual tools that automatically detect various types of harmful content combined with evaluations by human moderators under fair conditions [10]. Moreover, statistics about the diversity of the content (e.g. showing the regions from where the content originates) should be more common.

Improve interpretability to evaluate downstream effects on models. More transparency of filtered Common Crawl datasets should be complemented with better means to evaluate the effects of individual datasets on model behavior. Understanding how filtered Common Crawl versions are driving harmful outcomes is crucial not just for transparency, but also for fairness and accountability. Standardization efforts for evaluation frameworks, like EleutherAI's

Language Model Evaluation Harness [15] are important steps in this direction, but technical advancements in AI interpretability are necessary as well.

8 CONCLUSION

Despite Common Crawl's popularity for LLM development, its implications for fair, accountable, and transparent LLM research and development have garnered relatively little discussion so far. In this paper, we described how LLM builders rely on Common Crawl, and examined its mission and the crawling process in-depth to highlight the biases, advantages, and shortcomings of its data. We showed that Common Crawl has contributed to making LLM research and development more transparent and auditable, but that it at the same time is a problematic source for LLM training data that should be used with care. Long term, it would be beneficial to the field to see more values-driven intermediaries that filter Common Crawl in transparent and accountable ways. Reducing the reliance on Common Crawl (and similar sources) in favor of more curated datasets could lead to better outcomes for reducing risks of bias and harms in generative AI.

Future research could further investigate Common Crawl's historical evolution. As [26] shows, Common Crawl's approach to crawling the web has changed significantly over time. If LLM builders include older data that reaches back far enough, this history is important to further evaluate the implications for the resulting model. In addition, more socio-technical analysis of the filtering process for popular Common Crawl versions as in [29] would help illuminate why certain filtering techniques are popular, and what could be changed to help align these techniques better with fair, accountable, and transparent LLM research and development.

9 POSITIONALITY STATEMENT

The author has a background in media and journalism studies and works for a nonprofit organization based in the US. He is white, uses he/him pronouns, and is based in Europe. He mostly conducts qualitative research (interviews, ethnography), but also has experience in quantitative and computational methods. His work is committed to accountability, inclusion, and fairness.

REFERENCES

[1] Emily M. Bender, 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 (FAccT '21)*, March 01, 2021. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[2] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. https://doi.org/10.48550/arXiv.2304.01373

[3] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The Values Encoded in Machine Learning Research. *ArXiv210615590 Cs* (June 2021). Retrieved November 26, 2021 from http://arxiv.org/abs/2106.15590

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

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv211001963 Cs* (October 2021). Retrieved March 17, 2022 from http://arxiv.org/abs/2110.01963

[6] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological., Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf and Kenneth J. Sher (eds.). American Psychological Association, Washington, 57–71. https://doi.org/10.1037/13620-004

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

[8] Common Crawl Foundation. Our Mission. Retrieved November 2, 2023 from https://commoncrawl.org/mission

[9] Alejandro Cremades. 2019. Gil Elbaz On Google Acquiring His Company And Turning It Into A $15 Billion Business. *Alejandro Cremades*. Retrieved October 17, 2023 from https://alejandrocremades.com/gil-elbaz-on-google-acquiring-his-company-and-turning-it-into-a-15-billion-business/

[10] Jenny L. Davis, Apryl Williams, and Michael W. Yang. 2021. Algorithmic reparation. *Big Data Soc.* 8, 2 (July 2021). https://doi.org/10.1177/20539517211044808

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

[12] Jesse Dodge, 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. https://doi.org/10.48550/arXiv.2104.08758

[13] Leo Gao. 2021. An Empirical Exploration in Quality Filtering of Text Data. https://doi.org/10.48550/arXiv.2109.00698

[14] 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. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. https://doi.org/10.48550/arXiv.2101.00027

[15] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.10256836

[16] Brian Hayes. 2015. Crawling toward a Wiser Web. *Am. Sci.* 103, 3 (2015), 184. https://doi.org/10.1511/2015.114.184

[17] 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. 2023. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. https://doi.org/10.48550/arXiv.2303.03915

[18] Kalev Leetaru. 2017. Common Crawl And Unlocking Web Archives For Research. *Forbes*. Retrieved February 6, 2023 from https://www.forbes.com/sites/kalevleetaru/2017/09/28/common-crawl-and-unlocking-web-archives-for-research/

[19] Alexandra Sasha Luccioni and Joseph D. Viviano. 2021. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. https://doi.org/10.48550/arXiv.2105.02732

[20] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media Soc.* 19, 3 (March 2017), 329–346. https://doi.org/10.1177/1461444815608807

[21] McKinsey. 2023. What is generative AI? Retrieved October 13, 2023 from https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

[22] Sara Morrison. 2023. The tricky truth about how generative AI uses your data. *Vox*. Retrieved December 4, 2023 from https://www.vox.com/technology/2023/7/27/23808499/ai-openai-google-meta-data-privacy-nope

[23] Luke Munn. 2022. *Countering the Cloud: Thinking With and Against Data Infrastructures* (1st ed.). Routledge, New York. https://doi.org/10.4324/9781003341185

[24] Sebastian Nagel. 2019. commoncrawl vs archive.org etc. *Common Crawl mailing list*. Retrieved October 25, 2023 from https://groups.google.com/g/common-crawl/c/RBFAn0o55cY/m/68qiLwZMBAAJ

[25] Sebastian Nagel. 2022. Questions about using Common Crawl for another Hugging Face project. *Common Crawl mailing list*. Retrieved October 25, 2023 from https://groups.google.com/g/common-crawl/c/BgPvP6HB2n0/m/P-Nw5YoJAQAJ

[26] Sebastian Nagel. 2023. Common Crawl: Data Collection and Use Cases for NLP. In *HPLT & NLPL Winter School on Large-Scale Language Modeling and Neural Machine Translation with Web Data*. Retrieved August 3, 2023 from http://nlpl.eu/skeikampen23/nagel.230206.pdf

[27] Safiya Umoja Noble. 2018. *Algorithms of oppression: how search engines reinforce racism.* New York university press, New York.

[28] Will Orr. 2023. 9 Ways To See A Dataset: Datasets as sociotechnical artifacts — The case of "Colossal Cleaned Common Crawl" (C4). Retrieved January 16, 2024 from https://knowingmachines.org/publications/9-ways-to-see/essays/c4

[29] Will Orr and Kate Crawford. 2023. The social construction of datasets: On the practices, processes and challenges of dataset creation for machine learning. (2023). https://doi.org/10.31235/osf.io/8c9uh

[30] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv.org*. Retrieved June 22, 2023 from https://arxiv.org/abs/2203.02155v1

[31] Trevor Owens. 2014. Machine Scale Analysis of Digital Collections: An Interview with Lisa Green of Common Crawl – Coffeehouse. Retrieved October 17, 2023 from https://coffeehouse.dataone.org/2014/01/29/machine-scale-analysis-of-digital-collections-an-interview-with-lisa-green-of-common-crawl/

[32] Guilherme Penedo, 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. https://doi.org/10.48550/arXiv.2306.01116

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)*, 2014. 1532–1543. Retrieved from http://www.aclweb.org/anthology/D14-1162

[34] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? https://doi.org/10.48550/arXiv.2006.16923

[35] Colin Raffel, 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. https://doi.org/10.48550/arXiv.1910.10683

[36] Rémi Rampin and Vicky Rampin. 2021. Taguette: open-source qualitative data analysis. *J. Open Source Softw.* 6, 68 (December 2021), 3522. https://doi.org/10.21105/joss.03522

[37] Bruce Rogers. 2014. Gil Elbaz Builds Factual To Be The World's Data Steward. *Forbes*. Retrieved October 17, 2023 from https://www.forbes.com/sites/brucerogers/2014/05/29/gil-elbaz-builds-factual-to-be-the-worlds-data-steward/

[38] Kevin Schaul, Szu Yu Chen, and Nitasha Tiku. 2023. Inside the secret list of websites that make AI like ChatGPT sound smart. *Washington Post*. Retrieved June 12, 2023 from https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[39] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. https://doi.org/10.1145/3476058

[40] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput. Surv.* 55, 13s (July 2023), 293:1-293:39. https://doi.org/10.1145/3588433

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

[43] James Vincent. 2023. OpenAI co-founder on company's past approach to openly sharing research: "We were wrong." *The Verge*. Retrieved April 26, 2024 from https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview

[44] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. https://doi.org/10.48550/arXiv.1911.00359

[45] Adrienne Williams, Milagros Miceli, and Timnit Gebru. 2022. The Exploited Labor Behind Artificial Intelligence. *Noema*. Retrieved January 20, 2023 from https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence

[46] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

https://doi.org/10.48550/arXiv.2211.05100

[47] Jennifer Zaino. 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*. Retrieved October 17, 2023 from 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/

[48] Meta Llama FAQs. *Meta Llama Troubleshooting & FAQ*. Retrieved April 26, 2024 from https://llama.meta.com/faq/