Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Practical Datasets for Analyzing LLM Corpora Derived from Common Crawl

Nick Hagar[1], **Jack Bandy**[2]

[1]Northwestern University
[2]Transylvania University
nicholas.hagar@northwestern.edu, jbandy@transy.edu

Abstract

Large language models (LLMs) rely heavily on web-derived training datasets, yet understanding how filtering and curation decisions affect these datasets remains challenging. This paper presents two complementary datasets designed to enable systematic analysis of LLM training data composition. The first dataset captures domain-level statistics across 96 Common Crawl snapshots, providing baseline data about web content distribution before filtering. The second dataset contains standardized URL information from three major LLM training corpora (C4, Falcon RefinedWeb, and CulturaX), allowing researchers to analyze how different filtering approaches affect content inclusion. By making these datasets publicly available in a consistent format, we aim to (1) facilitate research into training data composition, (2) enable systematic auditing of filtering effects, and (3) support more transparent approaches to dataset development. Our datasets can help researchers investigate questions related to content diversity, source representation, and the impact of different filtering decisions on training data composition. Overall, this work provides a foundation for understanding how curation choices shape the content that ultimately trains widely-deployed language models.

1 Introduction

The rise of large language models (LLMs) in machine learning has increased demand for high-quality digital text. In particular, models like GPT-4, Llama 3, and Gemini require enormous amounts of diverse training data (Achiam et al. 2023; Dubey et al. 2024; Team et al. 2023). Consequently, the the contents of this training data has been scrutinized both in academic research (Dodge et al. 2021; Bandy and Vincent 2021; Birhane, Prabhu, and Kahembwe 2021) and the popular press (Queen 2023; Seetharaman 2024).

Often these LLM training datasets do not collect text directly from the web, but rather filter and re-combine existing data sources such as Common Crawl. However, beyond high-level model performance, the actual effect of the many filters, re-combinations, and refinements applied to these datasets remains an open question. Empirical evidence has demonstrated ample opportunities for bias to arise in large text corpora, intentionally or not (Gururangan et al. 2022), as training data tends to lack documentation and presents homogeneous, biased perspectives (Bender et al. 2021). When open-source datasets simply recombine an already biased dataset toward some measure of LLM data quality, without considering aspects like homogeneity, diversity, or factuality, they risk recreating biases and other issues in their training data.

To understand these potential biases and enable systematic investigation of training data composition, researchers need tools and frameworks for analyzing large-scale text collections. Currently, dataset documentation often focuses on technical metrics and high-level statistics, leaving important questions about content characteristics, source diversity, and representativeness unexplored. This gap is particularly notable given growing concerns about consent and representation in training data (Longpre et al. 2024a), alongside questions about how filtering and quality metrics affect the diversity of perspectives included.

This paper presents a framework and collection of datasets for investigating the composition of LLM training data, with a particular focus on understanding how filtering and recombination affect content representation. We first develop a pipeline to collect aggregate data about the domains present in 96 Common Crawl snapshots. These snapshots overlap with the data collection window leveraged by a number of popular LLM training datasets derived from Common Crawl, thus allowing researchers to directly analyze how filtering decisions affect dataset compositions. We also collect and compile URL information from three major LLM datasets, in a consistent format, as a resource for conducting comparison analyses in practice. Finally, we demonstrate a potential application of these datasets, with an examination of New York Times articles included in C4 and corresponding Common Crawl data.

These datasets provide a foundation for systematic investigation of LLM training data composition. The combination of longitudinal Common Crawl domain data and standardized URL information from major LLM datasets enables researchers to analyze filtering effects, track source inclusion patterns, and examine how different approaches to dataset curation affect content representation. By making these resources publicly available in a consistent, documented format, we aim to facilitate broader research into training data composition and support more transparent ap-

proaches to dataset development.

2 Background

The growing prominence of large language models has sparked increased scrutiny of their training data, particularly regarding issues of quality, bias, and representation. As models scale to process trillions of tokens, understanding the composition and characteristics of these massive training datasets becomes increasingly critical. Yet despite their importance, LLM training datasets often lack comprehensive documentation about filtering decisions, content representation, and potential biases. This opacity poses challenges for researchers attempting to analyze how different approaches to dataset curation affect model performance and fairness. Recent work has highlighted various concerns about web-scraped training data, from issues of consent and copyright to questions about data quality and demographic representation. However, systematically investigating these issues requires tools and frameworks that can efficiently analyze large-scale text collections at both broad and granular levels.

LLM Training Data Challenges

The task of training large language models presents a number of data-related challenges, many of which have yet to be fully addressed. At the most basic level, LLMs require a vast amount of training data in order to perform well on various benchmarks (Achiam et al. 2023). This has forced dataset curators to operate at ever-increasing scales to collect, filter, and validate training data. As Sambasivan et al. (2021) observed, this process of data curation has been an undervalued aspect of machine learning research, leading to many practices that may not be optimal for large-scale models.

One area of concern is related to legal and ethical considerations about different types of content in LLM training data. For example, researchers have found that models can memorize and reproduce sensitive personal information contained in training data (Carlini et al. 2021). Many LLM training datasets contain material from authors who did not give clear permission or consent to use their material for the purposes of training LLMs. As the landscape evolves in terms of copyright and fair use, the ability to identify specific content used to train specific models will be critically important.

Another common practice in LLM training data is to derive datasets from existing datasets, rather than directly designing and creating new training material. Major training datasets such as C4, Falcon RefinedWeb, and FineWeb all derive from Common Crawl—a free, Internet-scale database of web pages available through a flexible limited license agreement.[1] Some datasets, such as The Pile (Gao et al. 2020) and Dolma (Soldaini et al. 2024), include multiple sources and use Common Crawl as a main source of content alongside other inputs. As these datasets are derived from vast, unfiltered sources, they may be subject to a number of quality issues related to data freshness, diversity, bias, and more.

[1]https://commoncrawl.org/terms-of-use

Quality in LLM Training Data

The *quality* of LLM training data is key to model performance—in many cases, more so than the quantity of data used (Longpre et al. 2024c; Gunasekar et al. 2023). This creates a tension, wherein model trainers must strike a balance between collecting enough data for training an LLM to generate coherent outputs, while also ensuring that the included data are of high enough quality to support LLM performance across domains.

The first step in reaching this balance is defining what high-quality training data look like. Such definitions vary greatly, and tend to focus on aspects of desired model performance. At a high level, researchers deploy heuristics of quality—desirable training data might be written and edited by humans (Albalak et al. 2024), or it might be synthetic data that follows a certain form or structure (Gunasekar et al. 2023). If a model is geared toward a specific domain (e.g., programming), quality training data may focus on well-crafted examples of a specific type of text (e.g., source code — Hui et al. 2024). These varied definitions of quality underscore that training data selection is inherently tied to the intended capabilities and applications of the resulting model. These intentions then inform the approach that model builders take to crafting pipelines that identify and retrieve high-quality text from large corpora in practice, via either heuristic approaches or machine learning models (Brown et al. 2020; Rae et al. 2022).

However, quality cannot be defined along one dimension, or measured and implemented with one single approach. Model builders must confront trade-offs among a range of desirable characteristics related to diversity, timeliness, and toxic content (Yu et al. 2024). These trade-offs often create direct tensions in the filtering process. For example, aggressively filtering for toxicity may disproportionately remove content about marginalized groups or social justice issues, thus reducing text diversity (Dodge et al. 2021). These kinds of considerations can vary drastically depending on the specific task at hand. Whether a model is used for hate speech identification (Yin et al. 2023), value judgments (You and Suh 2024), persuasion (Breum et al. 2024), content moderation (Kumar, AbuHashem, and Durumeric 2024), or other use cases, it is important for model builders to intentionally curate training data in a way that aligns with the desired use case.

The vast quantity of LLM training data creates a situation where dataset quality is not measured directly. Instead, dataset curators use some kind of proxy, either in relation to heuristic characteristics of the data itself that will impact the model (Brown et al. 2020), or in relation to some quantification of model performance. The latter approach relies on benchmarks, which, while useful at a high level, are often an imperfect representation of only some of a model's functions (Davis 2024; Reuel et al. 2024). As a result, the process of data curation for model training is subject to numerous researcher degrees of freedom. Researchers make decisions about which benchmarks will best capture model performance, which characteristics of text data will have the most direct relationship to those benchmarks, and how to operationalize filters on those characteristics. These decisions

Dataset	Common Crawl Range	Citation
C4 English	CC-MAIN-2019-18	(Raffel et al. 2020)
Falcon RefinedWeb	CC-MAIN-2013-20 – CC-MAIN-2023-06	(Penedo et al. 2023)
FineWeb	CC-MAIN-2013-20 – CC-MAIN-2024-18	(Penedo et al. 2024)
MADLAD-400	CC-MAIN-2013-20 – CC-MAIN-2022-27	(Kudugunta et al. 2023)
DCLM	CC-MAIN-2013-20 – CC-MAIN-2022-49	(Li et al. 2024)
Dolma	CC-MAIN-2020-05 – CC-MAIN-2023-06	(Soldaini et al. 2024)

Table 1: Common Crawl snapshot ranges used in a sample of major language model training datasets

ultimately determine the contents of the training corpus. And while these decisions rely on expert assessment, LLM training datasets are typically vast and heterogeneous, limiting what humans can directly evaluate. The process raises tensions across domains, genres, and quality considerations, as well as opportunities for various kinds of bias and other undesirable characteristics to propagate in curated datasets.

Ultimately, heuristic decisions made during dataset curation can include data which is not aligned with the desired model performance (or, conversely exclude data which *is* aligned aligned with desired performance). The datasets described in this paper allow researchers to further explore this possibility, enabling more fine-grained analysis of the contents of LLM training corpora.

Dataset Documentation

Documenting the dataset construction process has become crucial for model builders and researchers. Building on the notion of "technical debt" (Zazworka et al. 2013), Bender et al. (2021) identified "documentation debt" as a critical issue in developing and maintaining large language models. This documentation debt manifests in two key ways: First, in the absence of clear records about what filtering decisions were made and why, and second, in the lack of systematic analysis regarding how these decisions affect the resulting dataset's composition and characteristics.

The implications of documentation debt are particularly significant for large training corpora. When datasets are sparsely documented, it becomes difficult to understand potential biases in model performance or navigate legal and copyright requirements. For example, dataset documentation can help characterize different types of bias—in medical imaging datasets, documentation has revealed how disparities can stem from prevalence, presentation, and annotation practices (Jones et al. 2024). Similarly, as the legal landscape around training data evolves, robust documentation of data provenance and filtering decisions will likely become increasingly important for compliance.

Several approaches have emerged to address these documentation challenges. Some researchers have created retrospective documentation for widely-used datasets, such as identifying qualitative "genealogies" for datasets like ImageNet (Denton et al. 2021), or tracing origins and transformations of datasets throughout their "life cycles" (Koch et al. 2021). These genealogies can reveal how initial curation decisions propagate through derivative datasets, potentially amplifying certain biases or gaps in coverage.

Datasheets offer another strategy for standardizing dataset documentation (Gebru et al. 2021; Hutchinson et al. 2021). By reporting key characteristics such as funding sources, acquisition processes, and descriptive statistics about the dataset's composition, datasheets support transparency and reproducibility in machine learning research. They also provide a framework for documenting the specific filtering and transformation steps applied during dataset creation, helping future users understand how the dataset's contents may have shifted from its source material. However, even with these documentation frameworks, capturing the full impact of curation decisions remains challenging. The scale of these datasets makes comprehensive documentation of filtering effects impractical, and the interaction between different filtering steps can be complex. Moreover, while documentation can reveal what steps were taken, it may not fully capture how these steps affect specific domains or types of content. The datasets described in this paper can help address these questions.

Data Sources and Datasets

One characteristic that unites many LLM training corpora is the recycling of raw data to create derivative datasets. In particular, many curated corpora use the Common Crawl—an Internet-scale, freely-available database of web pages—as their source of raw data. The RefinedWeb dataset, used to train the Falcon series of LLMs, is made up entirely of data which was cleaned and filtered from the Common Crawl (Penedo et al. 2023). Similarly, the C4 dataset is derived from one month (April 2019) of Common Crawl web pages, and FineWeb is built from 96 snapshots of Common Crawl data (Raffel et al. 2020; Penedo et al. 2024).

Some training datasets combine a variety of sources—The Pile, for example, contains ebooks, Wikipedia pages, medical documents, and other domain- or medium-specific text alongside a substantial amount of web data from the Common Crawl (Gao et al. 2020). But often, even the datasets that rely on a broader mix of data ultimately trace back a substantial number of documents to the Common Crawl. Dolma, for example, is a dataset that relies on Common Crawl directly. It also incorporates data from the C4 and RefinedWeb corpora, two datasets that themselves are entirely made up of Common Crawl text (Soldaini et al. 2024).[2]

[2]For a summary of the data sources contained in Dolma 1.7, see https://allenai.org/blog/olmo-1-7-7b-a-24-point-improvement-on-mmlu-92b43f7d269d

Given that large pretraining datasets largely rely on common sources of web text, their differentiation comes from the curation decisions made by dataset creators. Content mix, for example, can vary: While one dataset may rely entirely on web text, another might incorporate a blend of preprints, coding examples, and ebooks, among other formats, alongside its large dump of Common Crawl pages (Soldaini et al. 2024; Gao et al. 2020). The relative weight that various sources have in the mix of training data can affect the extent to which they influence model behavior.

Another point of differentiation involves the unique filtering, deduplication, and transformation approaches used to construct each dataset. C4 uses a range of heuristic rules (e.g., removing lines without a terminal punctuation mark, discarding pages with fewer than three sentences) as well as a classifier to identify English-language documents (Raffel et al. 2020). FineWeb adopts the C4 filters, as well as additional URL filtering and heuristics for quality (Penedo et al. 2024). Falcon RefinedWeb uses a similar mix of URL filters and document-level heuristics (Penedo et al. 2023).

These interconnected challenges around training data quality, documentation, and analysis point to a clear need for systematic approaches to understanding LLM training datasets. While existing documentation efforts have improved transparency around dataset construction, researchers still lack efficient tools for analyzing how filtering decisions and quality metrics affect content representation across massive web-scale corpora. The datasets presented in this paper aim to address this gap by providing standardized, queryable information about both raw web crawl data and the filtered datasets derived from it. By enabling direct comparison between Common Crawl snapshots and major training corpora, these resources can support investigation of critical questions about content diversity, source representation, and the effects of different filtering approaches. This type of systematic analysis is essential not only for understanding existing datasets but also for developing more transparent and intentional approaches to dataset curation as language models continue to grow in scale and importance.

3 Dataset Generation Methods

While LLM training datasets are primarily valued for their text content, understanding their composition requires analyzing metadata like URLs and domains. This source information provides crucial insights into content origin, diversity, and potential biases in training data. However, comprehensive analysis of these datasets presents significant technical challenges due to their scale. Common Crawl snapshots contain billions of records, while derived training datasets often include hundreds of millions of documents. The sheer volume of text data makes direct analysis computationally intensive and often impractical.

To make these datasets more amenable to systematic audit and analysis, we extract and compile URL and domain information from both raw Common Crawl data and derived training datasets. Our approach uses SQL queries, providing an efficient and replicable method for processing data at this scale.[3] We focus on two complementary data collection efforts that together enable investigation of how filtering decisions affect training data composition.

Dataset	Size (GB)
Common Crawl total (96 snapshots)	312GB
Common Crawl median snapshot	3GB
CulturaX URLs	395GB
Falcon RefinedWeb URLs	51GB
C4 English URLs	20GB

Table 2: Dataset sizes in compressed gigabytes (GB)

The first collection effort focuses on Common Crawl data. Common Crawl is structured as a series of periodic snapshots, each representing a complete web crawl. These snapshots serve as the foundation for many LLM training datasets, which typically combine and filter content from multiple crawls. Each snapshot is also extremely large, making them difficult to process for downstream analysis—the December 2024 crawl, for example, contains 394 TiB of uncompressed data, encompassing 2.64 billion web pages (Nagel 2024).

Our goal in processing these snapshots is to allow straightforward querying of record identifiers, separated from the network bandwidth and compute considerations required to process the full text corpora. To do so, we use Amazon's Athena query engine to access Common Crawl's publicly available index data on S3. Because of the scale of each snapshot, we aggregate records at the domain level, computing the total number of pages included from each source domain. We store the resulting domain-level statistics for each snapshot as separate datasets on HuggingFace[4], providing a baseline for analyzing how subsequent filtering affects source distribution. We collect data from 96 Common Crawl snapshots, covering the time period leveraged by many prominent training datasets (see Table 1).

Our second collection effort focuses on extracting record identifiers, in the form of URLs, from widely-used training corpora that derive text from Common Crawl. We select training corpora to include by searching HuggingFace for large-scale, curated collections that have been used to train widely-deployed language models. For this initial implementation of our collection approach, we focus on three datasets: the English variant of C4 (Raffel et al. 2020), Falcon RefinedWeb (used to train the Falcon family of LLMs—Almazrouei et al. 2023; Penedo et al. 2023), and multilingual corpus CulturaX (Nguyen et al. 2024).

For each dataset, we collect the complete set of URLs from their repositories on Hugging Face using the DuckDB database engine (Mühleisen and Raasveldt 2025). Because these corpora are smaller in scale than the Common Crawl data they rely on, we preserve the raw URLs from these derived datasets. This granular data enables researchers to both

[3] https://github.com/NHagar/cc-genealogy

[4] https://huggingface.co/collections/nhagar/cc-domain-counts-67645a737b7a300ad3ab539f

compute domain-level statistics and conduct more detailed analyses of content inclusion patterns.

All collected data are stored in standardized formats on HuggingFace[5], with unique DOIs for each Common Crawl snapshot and derived dataset. By providing efficient access to source information from both raw web crawls and filtered training datasets, these resources enable systematic investigation of how dataset curation decisions affect content representation in LLM training data.

To the extent possible given the computation requirements of processing data at scale, we host these datasets with minimal preprocessing to allow for open-ended analysis across a range of interest areas. While the resulting datasets are still large (see Table 2), we hope that separating record identifiers, in the form of URLs and domains, from the massive corpora of text used to train LLMs makes auditing, describing, and analyzing the text underpinning state-of-the-art models more tractable for a broader range of researchers.

4 Dataset Description

Our collection comprises two datasets that together enable detailed analysis of LLM training data composition. The first captures 582 million unique domains across 96 Common Crawl snapshots, encompassing 310 billion total URLs. The second contains approximately 8 billion URLs from three prominent LLM training corpora, detailed in Table 4.

Common Crawl The Common Crawl snapshots reveal the raw material that training datasets are built from. Like many web-scale collections, these snapshots show a heavy-tailed distribution of content, but with interesting heterogeneity. For example, in the most recent snapshot in our collection (CC-MAIN-2024-18), the top 10 domains only make up 0.06% of records, but they contribute a median 203,000 URLs each. This is far higher than the median domain overall, which only contributes 4 records. The 10 most prevalent sites for this snapshot (Table 3) highlight the range of sources included in the crawl: government resources, academic publications, user-generated content, and even e-commerce all contribute a large number of pages. This distribution gives us a window into Common Crawl's crawling patterns and provides crucial context for understanding subsequent filtering decisions.

Training corpora Our derived datasets let researchers directly compare this raw web data against the cleaned versions used in model training. The filtered English version of the C4 corpus, for example, contains 365,000,500 URLs, 12% of what appears in its source Common Crawl snapshot. These comparisons can reveal how curation reshapes content distribution. For example, while nytimes.com makes up 0.02% (640,000 records) of the relevant Common Crawl snapshot, this share grows to 0.05% in C4 (170,000 records). By querying the index of this Common Crawl snapshot, we can also retrieve the underlying nytimes.com URLs. This reveals a shift in distribution across site sections in the curated dataset, when looking at all sections that make up at least 1% of New York Times pages in either dataset: The curated C4 dataset relies less on archival coverage, while increasing the proportion of articles from sections like U.S. and world news, sports, and New York regional coverage (Figure 1).

[5]https://huggingface.co/collections/nhagar/llm-training-urls-67645a94115fa772cb1f89f8

Domain	URLs	Percentage (%)
www.ncbi.nlm.nih.gov	326,881	0.10
social-plugins.line.me	258,385	0.08
sso.sagepub.com	240,557	0.07
pubmed.ncbi.nlm.nih.gov	221,504	0.07
www.youtube.com	210,548	0.06
access.line.me	194,863	0.06
sites.google.com	185,125	0.06
learn.microsoft.com	179,461	0.05
dx.doi.org	172,390	0.05
us.vestiairecollective.com	171,577	0.05

Table 3: Top 10 domains by number of URLs in the CC-MAIN-2024-18 snapshot

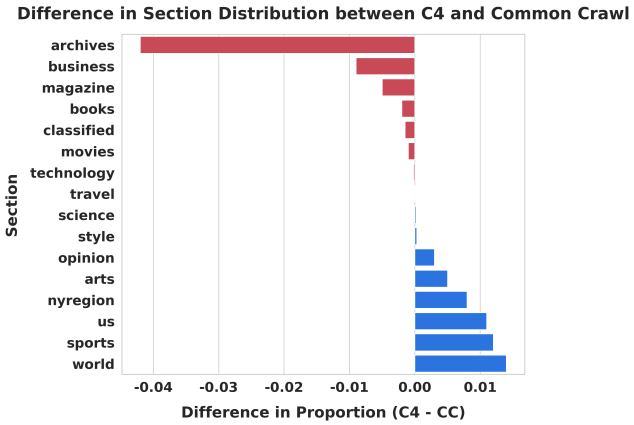

Figure 1: Differences in content distribution across sections between C4 and Common Crawl datasets. Positive values (blue) indicate higher representation in C4, while negative values (red) show sections more prevalent in Common Crawl. Notable differences include C4's underrepresentation of archive content and over representation of sports and world news.

FAIR Principles Our datasets follow FAIR principles described by (FORCE11 2020). The data are **Findable** through persistent URLs Hugging Face, including through organized collections. Each collection has a descriptive title and is linked in our paper. The data are **Accessible** through Hugging Face interfaces and APIs, such that researchers can access the data with common tools. In terms of **Interoperability**, the datasets use standardized formats and clear schema. Finally, the datasets are **Reusable** given the documentation provided for this paper, standardized data formats, and citation information for source datasets.

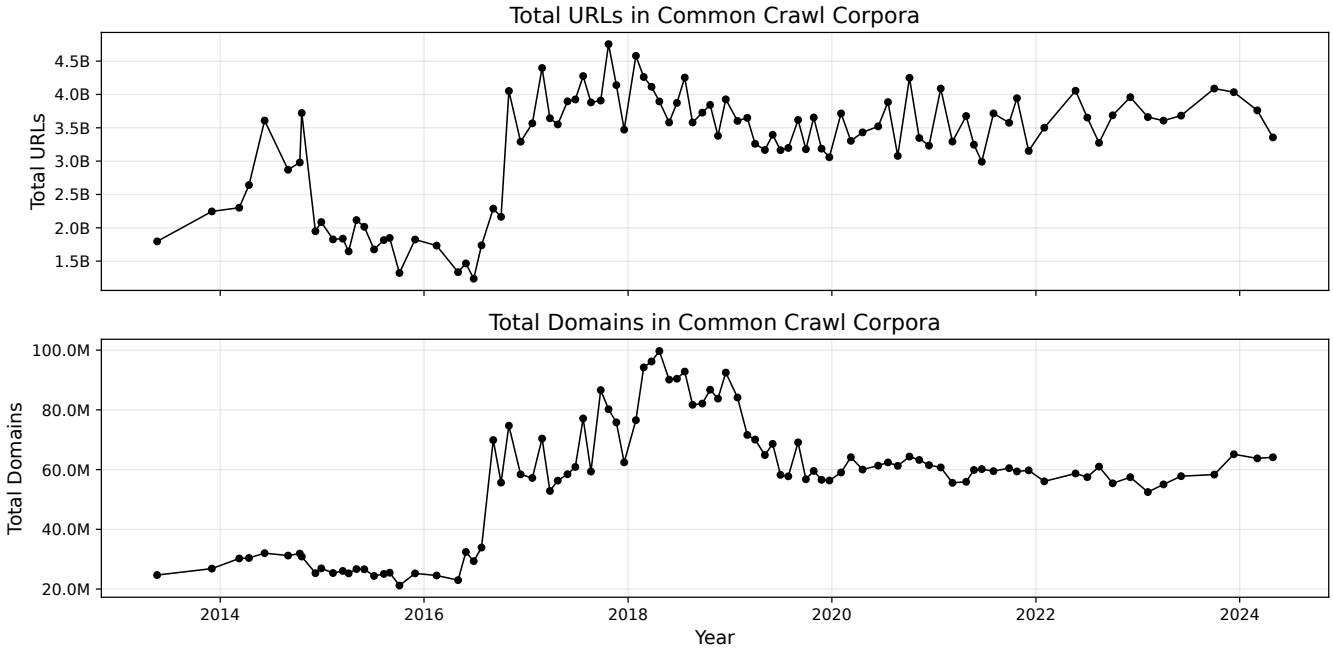

Figure 2: Visualization of Common Crawl snapshot statistics from 2013 to 2024, showing total URLs (top) and unique domains (bottom) in each snapshot.

Dataset	Records	Notable Model(s)	License
CulturaX	7,184,124,703	Stable LM 2 family (Bellagente et al. 2024)	odc-by/cc0-1.0
Falcon RefinedWeb	968,000,015	Falcon family (Almazrouei et al. 2023)	odc-by
C4 (English)	365,233,500	LLaMA (Touvron et al. 2023)	odc-by

Table 4: Record counts, language models trained, and licenses for Common Crawl-derived corpora. These datasets represent influential examples of filtered and curated web text that has been successfully deployed in training large language models.

5 Discussion

Limitations

The datasets described in this paper are subject to a number of limitations. Common Crawl summary statistics are at the domain (host name) level, so some analyses may need to use raw index files from Common Crawl to analyze data at the URL level.

While the datasets described in this paper are intended for analytical and research purpose, there is also potential for misuse of the data. While we do not include the text associated with the URLs in the dataset, URLs themselves can sometimes be used to reconstruct private information. Furthermore, the list of URLs could be used to scrape content from websites without regard for robots.txt files and other mechanisms for consent. In our view, the risks associated with publishing the datasets described in this paper are no greater than the risks associated with the existing published datasets.

Future Work

We anticipate the datasets in this paper can support a wide range of future research to deepen our understanding of web-based datasets. The structure of these datasets is especially well-suited for comparing different datasets or analyzing the evolution of a single dataset, as well as several other areas of research.

The broad, looming question about LLM training datasets is the extent to which they capture a useful sample of text data available on the internet. In the context of large-scale social media data, researchers regularly analyzed sampled datasets to see if they captured a valid representation of the entire platform (Pfeffer et al. 2023). Similarly, while LLM training datasets may not want a statistically representative sample of the entire internet, researchers and practitioners would benefit from further analysis regarding how these sampled datasets compare to the entire internet, for example, by comparing it to the Common Crawl baseline described in this paper.

The potential for comparing different LLM training datasets may also be especially fruitful. Although dataset documentation efforts have improved in recent years, these efforts tend to focus on individual datasets. By analyzing the distribution and concentration of different web domains in these datasets, researchers can identify distribution patterns

and potentially determine the curation decisions that affect those distributions. Ideally, these findings would lead to more fruitful, evidence-based practices in large-scale dataset curation.

More granular analyses could focus on the different types of domains represented in the datasets (e.g. news, legal, academic, social, etc.), or even the specific topics in the datasets. Such analyses could determine topical gaps in a dataset—for example, a model trained for code generation would ideally be trained on a wide range of programming languages and paradigms. A topical analysis would reveal how different languages and paradigms are represented in the training data.

Another compelling research direction involves analyzing the role of news content in these datasets. Because text from news publishers follows journalistic standards of structure, factuality, and more, "high-quality" datasets may disproportionately rely on text from news publishers (Gururangan et al. 2022; Longpre et al. 2024b). Future work may analyze the representation of different types of news publishers in these training datasets, from major national outlets to local news sources.

Because news articles are often written for (or about) specific locations, geographic analysis presents another compelling area for future work. Researchers might label URLs or domains with their associated location to analyze geographic representation in these datasets. By identifying over-represented or under-represented locations, such analyses may help future dataset curators improve geographic representation (Thelwall and Vaughan 2004).

Related to geographic representation is the question of linguistic representation. Datasets that include multiple languages could be analyzed to validate representation and compare text quality in different languages. As with the example of programming languages, a multi-lingual model may have gaps in training data that diminish performance in specific languages.

Finally, alongside comparative work and analyses focused on topical representation, news, geographic distributions, and linguistic representation, future work could trace the evolution of training data over time. Such work could follow the example of Longpre et al. (2024a), which showed that a growing proportion of web domains are explicitly opting out of sharing their data for LLM training. By tracking how data passes from Common Crawl into derivative datasets and even different versions of those datasets, researchers and practitioners can gain a deeper understanding of these vast datasets.

In addition to advancing our understanding of existing datasets, this future work can improve the development of future datasets. By identifying existing patterns and gaps, curators can make evidence-based decisions about what to include in LLM training datasets in order to support desired performance.

6 Conclusion

This paper presents two complementary datasets designed to enable systematic analysis of LLM training data composition and curation. By providing domain-level statistics across 96 Common Crawl snapshots alongside standardized URL information from major training corpora, these resources create new opportunities for investigating how filtering decisions shape the content that ultimately trains widely-deployed language models. The datasets support a range of critical analyses, from examining source diversity and representation to understanding how different quality metrics and filtering approaches affect content inclusion. As concerns about training data consent, quality, and bias continue to grow, the ability to systematically analyze dataset composition becomes increasingly vital. These datasets provide a foundation for more transparent and intentional approaches to dataset curation, enabling researchers to investigate how different filtering decisions impact content representation and potentially influence model behavior. Looking ahead, we anticipate these resources will support crucial research into training data composition, help identify potential gaps or biases in existing datasets, and inform the development of more robust and well-documented approaches to creating training corpora for future language models.

Code — https://github.com/NHagar/cc-genealogy

Datasets (Common Crawl) — https://huggingface.co/collections/nhagar/cc-domain-counts-67645a737b7a300ad3ab539f

Datasets (LLM Corpora) — https://huggingface.co/collections/nhagar/llm-training-urls-67645a94115fa772cb1f89f8

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report.

Albalak, A.; Elazar, Y.; Xie, S. M.; Longpre, S.; Lambert, N.; Wang, X.; Muennighoff, N.; Hou, B.; Pan, L.; Jeong, H.; Raffel, C.; Chang, S.; Hashimoto, T.; and Wang, W. Y. 2024. A Survey on Data Selection for Language Models. ArXiv:2402.16827.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, É.; Hesslow, D.; Launay, J.; Malartic, Q.; Mazzotta, D.; Noune, B.; Pannier, B.; and Penedo, G. 2023. The Falcon Series of Open Language Models. ArXiv:2311.16867 [cs].

Bandy, J.; and Vincent, N. 2021. Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Bellagente, M.; Tow, J.; Mahan, D.; Phung, D.; Zhuravinskyi, M.; Adithyan, R.; Baicoianu, J.; Brooks, B.; Cooper, N.; Datta, A.; Lee, M.; Mostaque, E.; Pieler, M.; Pinnaparju, N.; Rocha, P.; Saini, H.; Teufel, H.; Zanichelli, N.; and Riquelme, C. 2024. Stable LM 2 1.6B Technical Report.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 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*, 610–623.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 152–163.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Carlini, N.; Tramèr, F.; Brown, T.; Song, D.; and Oprea, A. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium*.

Davis, E. 2024. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Computing Surveys*, 56(4): 1–41.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; and Nicole, H. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 20539517211035955.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: October 12, 2024.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. ArXiv:2101.00027 [cs].

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Giorno, A. D.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; Rosa, G. d.; Saarikivi, O.; Salim, A.; Shah, S.; Behl, H. S.; Wang, X.; Bubeck, S.; Eldan, R.; Kalai, A. T.; Lee, Y. T.; and Li, Y. 2023. Textbooks Are All You Need. ArXiv:2306.11644.

Gururangan, S.; Card, D.; Dreier, S.; Gade, E.; Wang, L.; Wang, Z.; Zettlemoyer, L.; and Smith, N. A. 2022. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2562–2580. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Hui, B.; Yang, J.; Cui, Z.; Yang, J.; Liu, D.; Zhang, L.; Liu, T.; Zhang, J.; Yu, B.; Lu, K.; Dang, K.; Fan, Y.; Zhang, Y.; Yang, A.; Men, R.; Huang, F.; Zheng, B.; Miao, Y.; Quan, S.; Feng, Y.; Ren, X.; Ren, X.; Zhou, J.; and Lin, J. 2024. Qwen2.5-Coder Technical Report. ArXiv:2409.12186 [cs].

Hutchinson, B.; Smart, A.; Hanna, A.; Denton, E.; Greer, C.; Kjartansson, O.; Barnes, P.; and Mitchell, M. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575.

Jones, C.; Castro, D. C.; De Sousa Ribeiro, F.; Oktay, O.; McCradden, M.; and Glocker, B. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2): 138–146.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Kudugunta, S.; Caswell, I.; Zhang, B.; Garcia, X.; Xin, D.; Kusupati, A.; Stella, R.; Bapna, A.; and Firat, O. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 67284–67296. Curran Associates, Inc.

Kumar, D.; AbuHashem, Y. A.; and Durumeric, Z. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 865–878.

Li, J.; Fang, A.; Smyrnis, G.; Ivgi, M.; Jordan, M.; Gadre, S.; Bansal, H.; Guha, E.; Keh, S.; Arora, K.; Garg, S.; Xin, R.; Muennighoff, N.; Heckel, R.; Mercat, J.; Chen, M.; Gururangan, S.; Wortsman, M.; Albalak, A.; Bitton, Y.; Nezhurina, M.; Abbas, A.; Hsieh, C.-Y.; Ghosh, D.; Gardner, J.; Kilian, M.; Zhang, H.; Shao, R.; Pratt, S.; Sanyal, S.; Ilharco, G.; Daras, G.; Marathe, K.; Gokaslan, A.; Zhang, J.; Chandu, K.; Nguyen, T.; Vasiljevic, I.; Kakade, S.; Song, S.; Sanghavi, S.; Faghri, F.; Oh, S.; Zettlemoyer, L.; Lo, K.; El-Nouby, A.; Pouransari, H.; Toshev, A.; Wang, S.; Groeneveld, D.; Soldaini, L.; Koh, P. W.; Jitsev, J.; Kollar, T.; Dimakis, A. G.; Carmon, Y.; Dave, A.; Schmidt, L.; and Shankar, V. 2024. DataComp-LM: In search of the next generation of training sets for language models. arXiv:2406.11794.

Longpre, S.; Mahari, R.; Lee, A.; Lund, C.; Oderinwale, H.; Brannon, W.; Saxena, N.; Obeng-Marnu, N.; South, T.; Hunter, C.; et al. 2024a. Consent in crisis: The rapid decline of the ai data commons. In *NEURIPS*.

Longpre, S.; Singh, N.; Cherep, M.; Tiwary, K.; Materzynska, J.; Brannon, W.; Mahari, R.; Dey, M.; Hamdy, M.; Saxena, N.; Anis, A. M.; Alghamdi, E. A.; Chien, V. M.; Obeng-Marnu, N.; Yin, D.; Qian, K.; Li, Y.; Liang, M.; Dinh, A.; Mohanty, S.; Mataciunas, D.; South, T.; Zhang, J.; Lee, A. N.; Lund, C. S.; Klamm, C.; Sileo, D.; Misra, D.; Shippole, E.; Klyman, K.; Miranda, L. J.; Muennighoff, N.; Ye, S.; Kim, S.; Gupta, V.; Sharma, V.; Zhou, X.; Xiong, C.; Villa, L.; Biderman, S.; Pentland, A.; Hooker, S.; and Kabbara, J. 2024b. Bridging the Data Provenance Gap Across Text, Speech and Video. ArXiv:2412.17847 [cs].

Longpre, S.; Yauney, G.; Reif, E.; Lee, K.; Roberts, A.; Zoph, B.; Zhou, D.; Wei, J.; Robinson, K.; Mimno, D.; and Ippolito, D. 2024c. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3245–3276. Mexico City, Mexico: Association for Computational Linguistics.

Mühleisen, H.; and Raasveldt, M. 2025. *duckdb: DBI Package for the DuckDB Database Management System*. R package version 1.1.3.9029, https://github.com/duckdb/duckdb-r.

Nagel, S. 2024. Common Crawl - Blog - December 2024 Crawl Archive Now Available.

Nguyen, T.; Nguyen, C. V.; Lai, V. D.; Man, H.; Ngo, N. T.; Dernoncourt, F.; Rossi, R. A.; and Nguyen, T. H. 2024. CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 4226–4237. Torino, Italia: ELRA and ICCL.

Penedo, G.; Kydlíček, H.; allal, L. B.; Lozhkov, A.; Mitchell, M.; Raffel, C.; Werra, L. V.; and Wolf, T. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. ArXiv:2406.17557.

Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. ArXiv:2306.01116 [cs].

Pfeffer, J.; Mooseder, A.; Lasser, J.; Hammer, L.; Stritzel, O.; and Garcia, D. 2023. This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 720–729.

Queen, J. 2023. Sarah Silverman sues Meta, OpenAI for copyright infringement. *Reuters*.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; Rutherford, E.; Hennigan, T.; Menick, J.; Cassirer, A.; Powell, R.; Driessche, G. v. d.; Hendricks, L. A.; Rauh, M.; Huang, P.-S.; Glaese, A.; Welbl, J.; Dathathri, S.; Huang, S.; Uesato, J.; Mellor, J.; Higgins, I.; Creswell, A.; McAleese, N.; Wu, A.; Elsen, E.; Jayakumar, S.; Buchatskaya, E.; Budden, D.; Sutherland, E.; Simonyan, K.; Paganini, M.; Sifre, L.; Martens, L.; Li, X. L.; Kuncoro, A.; Nematzadeh, A.; Gribovskaya, E.; Donato, D.; Lazaridou, A.; Mensch, A.; Lespiau, J.-B.; Tsimpoukelli, M.; Grigorev, N.; Fritz, D.; Sottiaux, T.; Pajarskas, M.; Pohlen, T.; Gong, Z.; Toyama, D.; d'Autume, C. d. M.; Li, Y.; Terzi, T.; Mikulik, V.; Babuschkin, I.; Clark, A.; Casas, D. d. L.; Guy, A.; Jones, C.; Bradbury, J.; Johnson, M.; Hechtman, B.; Weidinger, L.; Gabriel, I.; Isaac, W.; Lockhart, E.; Osindero, S.; Rimell, L.; Dyer, C.; Vinyals, O.; Ayoub, K.; Stanway, J.; Bennett, L.; Hassabis, D.; Kavukcuoglu, K.; and Irving, G. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. ArXiv:2112.11446 [cs].

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. ArXiv:2411.12990 [cs].

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Seetharaman, D. 2024. For data-guzzling ai companies, the internet is too small. *The Wall Street Journal*.

Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Authur, R.; Bogin, B.; Chandu, K.; Dumas, J.; Elazar, Y.; Hofmann, V.; Jha, A. H.; Kumar, S.; Lucy, L.; Lyu, X.; Lambert, N.; Magnusson, I.; Morrison, J.; Muennighoff, N.; Naik, A.; Nam, C.; Peters, M. E.; Ravichander, A.; Richardson, K.; Shen, Z.; Strubell, E.; Subramani, N.; Tafjord, O.; Walsh, P.; Zettlemoyer, L.; Smith, N. A.; Hajishirzi, H.; Beltagy, I.; Groeneveld, D.; Dodge, J.; and Lo, K. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models.

Thelwall, M.; and Vaughan, L. 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26: 162–176.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Yin, W.; Agarwal, V.; Jiang, A.; Zubiaga, A.; and Sastry, N. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In

Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 902–913.

You, J.; and Suh, B. 2024. Evaluating and Improving Value Judgments in AI: A Scenario-Based Study on Large Language Models' Depiction of Social Conventions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1727–1739.

Yu, X.; Zhang, Z.; Niu, F.; Hu, X.; Xia, X.; and Grundy, J. 2024. What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 656–668. Sacramento CA USA: ACM. ISBN 9798400712487.

Zazworka, N.; Spínola, R. O.; Vetro', A.; Shull, F.; and Seaman, C. 2013. A case study on effectively identifying technical debt. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*, 42–47.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our dataset supports analysis of publicly available web crawl data and training datasets accessible through Common Crawl and Hugging Face.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. The abstract and introduction focus on the core contribution of creating efficient, queryable datasets for analyzing the composition of LLM training data based on domain and URL information.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Our methods section explains our choice of tools and the aggregation approach used for Common Crawl data.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. The paper describes the temporal nature of the Common Crawl snapshots, variations in crawl sizes over time, and the effect of domain-level aggregation.
   (e) Did you describe the limitations of your work? Yes, we include a limitations subsection in the Discussion section which highlights the scope of our analysis and temporal limitations of Common Crawl coverage.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. Potential negative impacts are described in the limitations section.
   (g) Did you discuss any potential misuse of your work? Yes. Potential misuse is described in the limitations section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We offer suggestions for using the data in the Future Work section, include methods details for reproducibility, and documentation in the data releases.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes. To our understanding, our paper and the release of these datasets conforms to the ethics guidelines.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes. Our datasets were created based on existing public datasets for pretraining LLMs. The paper includes citations and direct links to the source data.

(b) Did you mention the license of the assets? Yes, we include the licenses reported by the dataset creators in Table 4.

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? There are known instances of PII in Common Crawl and its derived datasets. We do not include raw text from Common Crawl, however, we preserve URL instances for accurate representation of record identifiers.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes. We include a subsection about FAIR principles in the main Dataset Description.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Datasheets are in-progress on HuggingFace.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA