Patterns

CellPress
OPEN ACCESS

Perspective

An archival perspective on pretraining data

Meera A. Desai,[1,*] Irene V. Pasquetto,[2] Abigail Z. Jacobs,[1] and Dallas Card[1]
[1]School of Information, University of Michigan, Ann Arbor, MI, USA
[2]College of Information Studies, University of Maryland, College Park, MD, USA
*Correspondence: madesai@umich.edu
https://doi.org/10.1016/j.patter.2024.100966

THE BIGGER PICTURE Large language models have become ubiquitous but depend crucially on the data on which they are trained. These pretraining datasets are themselves distinctive artifacts that are reused, built upon, and made legitimate beyond their role in shaping model outputs. We consider the similarities between pretraining datasets and archives: both are collections of diverse sociocultural materials that mediate knowledge production and thereby confer power to those who select, document, and control access to them. We discuss the limitations of current approaches to assembling pretraining datasets and ask whose voices are amplified or obscured? Who is harmed? Whose perspectives are taken up or assumed as the default? We highlight the need for more research on these datasets and the practices through which they are built and suggest possible paths forward, drawing on ideas from archival studies.

SUMMARY

Alongside an explosion in research and development related to large language models, there has been a concomitant rise in the creation of pretraining datasets—massive collections of text, typically scraped from the web. Drawing on the field of archival studies, we analyze pretraining datasets as informal archives—heterogeneous collections of diverse material that mediate access to knowledge. We use this framework to identify impacts of pretraining data creation and use beyond directly shaping model behavior and reveal how choices about what is included in pretraining data necessarily involve subjective decisions about values. In doing so, the archival perspective helps us identify opportunities for researchers who study the social impacts of technology to contribute to confronting the challenges and trade-offs that arise in creating pretraining datasets at this scale.

INTRODUCTION

The rise of large language models (LLMs) in recent years has been driven in large part by the creation of massive pretraining datasets, which are mostly composed of raw, unstructured text from the internet. While extensive research has studied the impacts of LLMs, including their potential for harm and associated mitigations,[1–6] researchers have devoted less critical attention to the data that enable the creation of these models. As we argue here, pretraining datasets are important not just for their influence on model behavior but as unique sociocultural collections, often with lasting impacts. Drawing on a long tradition in archival studies—a field that contends with assembling collections of sociocultural materials—we show how an archival perspective offers researchers who study the social impacts of technology useful insights into the inherent power of pretraining data, those who create it, and the practices that shape its development.[7,8]

Although pretraining datasets differ meaningfully from traditional archives, the two share some key features, as we will discuss below. In particular, like archives, pretraining datasets mediate access to material that is used for knowledge production. Here, we consider how pretraining datasets function as informal archives in order to theorize the importance of pretraining datasets and interrogate how these datasets are built and used. By focusing on the decisions made by dataset creators and the issues they prioritize, we underscore both the consequential nature of these decisions and the often unstated assumptions involved.

The rest of this paper is organized as follows: we first provide relevant background on archival studies and use it as a lens for conceptualizing pretraining datasets as collections of sociocultural material. We then consider the practices used in building pretraining datasets, focusing on three especially common concerns among LLM researchers. In particular, by drawing on the parallel between these practices and the archival practice of appraisal, we help to surface the values and assumptions underlying these areas of focus. Finally, we discuss the implications of an archival perspective for the creation of pretraining datasets and their impacts on knowledge production.

Table 1. Pretraining data possess similar attributes to traditional archives

	Traditional archives	Pretraining data for LLMs
What sociocultural materials are collected?	Documents, artifacts, etc.	Text data
What is known about the contents and origins of these materials?	Provenance of materials is documented through metadata; documentation is available to users through description and finding aids	Details may be unknown due to scale and lack of attention to provenance
What sensitive and contested materials are collected?	Private documents, human remains, colonial collections, etc.	Private information, copyrighted material, text that is demeaning to a social group, etc.
How do collections get authority and legitimacy?	Official status, use by historians, exclusive documents (e.g., private letters), connection with institutions such as universities or museums	Academic publications, authorial reputation, use in training models, connection with institutions such as universities or companies
What impact do collections have on knowledge production?	Mediate our understanding of the past, thereby impacting cultural memory and human identity	Impacts future practices: pretraining data curation, data reuse

Understanding the pretraining data as an informal archive can help reveal the implicit power involved in their creation and dissemination.

In some contexts, acting on these insights will require more effort and coordination (e.g., institutional support, external pressure, regulation, etc.) than others. Most pretraining datasets are created by a relatively small number of people in a variety of organizational contexts, including corporate (e.g., OpenAI,[9] Brown et al.[10]), academic (e.g., Gokaslan and Cohen[11]), and non-profit (e.g., Gao et al.,[12] Soldaini et al.[13]). In each of these contexts, creators are subjected to different kinds of incentives and pressures that may limit implementation of the practices we discuss (see Table 2 for examples). In all cases, we invite critical reflection and discussion among dataset creators. Most importantly, throughout the paper we argue for more critical attention on pretraining data from scholars who study the social impacts of technology, including from archival studies; information, computer, and social sciences; and related disciplines.

PRETRAINING DATA THROUGH THE LENS OF ARCHIVAL STUDIES

The modern notion of archives as public repositories of state documents emerged in the wake of the French Revolution.[14,15] Archivists were trained to abide by the principle of *respect des fonds*—dictating that material should be kept in its original order—and to record provenance– the context and history from which the material was drawn.[15,16] Facing ever-larger collections, archivists developed the practices of *appraisal* and *selection* to evaluate and choose materials for preservation, respectively.[14,17]

Over time, the field of archival studies has come to recognize the power-laden nature of archives[15] and their authority to legitimize the materials within them,[18] thereby shaping our knowledge and written history.[19–21] Since our understanding of the past mediates the formation of cultural memory and human identity,[15,22,23] critical scholarship has attended to the "silences of the archives"—that is, emphasizing what has not been included whether by chance, circumstance, or deliberate omission.[24,25] In addition to appraisal and selection, archival practices also influence how people think about and interact with the past via additional layers of infrastructure such as indices, guides, finding aids, and other forms of representation.[21,26,27] Archival scholarship thus confronts the challenges of assembling an archive while attending to the political stakes at hand.[28–30]

Although they are created for a different purpose than archives, pretraining datasets gather together a great variety of material in a stable, citable, and often named, repository (e.g., OpenWebText,[11] The Pile,[12] ROOTS,[31] etc.). Moreover, much like archivists, those who create pretraining datasets select, document, and mediate access to sociocultural materials that are used in knowledge production. An archival perspective suggests that we should attend to and interrogate the processes by which these datasets are created, how they are represented, and the effects that they have in the world[7,8] (see Table 1).

Pretraining data most obviously mediate knowledge production through their role in language models. Data choices determine not only the capabilities of a generative model but, more fundamentally, language affordances (e.g., English or multilingual) and information that is included. To the extent that people use LLMs as interfaces into history and culture, the selection of data shapes and constrains that experience.[32–34] In parallel to the appraisal of material for archives, those who appraise and select information for archives are exercising an important form of power, as we discuss in the following section. For archives, appraisal and selection involve the power to enable or limit what history can be written; with pretraining data, this power enables or constrains the potential of associated models.

However, even independent of models, the very act of including material in a pretraining corpus can lend legitimacy and authority to the use of these materials in knowledge production,[35] as is the case for traditional archives. Despite this, pretraining data often include legally and ethically contested data: most web-scale datasets are collected without consent.[36,37] Automated selection is virtually guaranteed to include sensitive and contested material, and questions of copyright and ownership are actively being litigated.[38–42]

Although it is not their primary function, pretraining datasets contribute to preserving historical data and thus have some power in mediating our access to information from the past. For example, because of the dynamic and unstable nature of the internet,[43,44] web-crawled pretraining datasets could end up being the only preserved copy of an edited or deleted webpage.

Table 2. Four examples of prominent pretraining datasets along with brief summaries of the appraisal and selection factors discussed in this paper and observations on downstream impacts

Dataset	WebText[88]	The Pile[12]	ROOTS Corpus[89]	Dolma[13]
Creators	Researchers at OpenAI ("capped profit" organization)	Grassroots effort, later incorporated as EleutherAI (non-profit organization)	BigScience research workshop, 1,000+ researchers from 60 countries	Researchers at the Allen Institute for AI (non-profit organization)
Language(s)	English	English	59 languages	English
Size, Tb	0.04	0.8	1.6	5.4 (v.1.6)
Dataset focus	Large and diverse dataset of high-quality text scraped directly from the web	High-quality text with greater domain diversity (e.g., web, books, academic, code, etc.)	Collaborative and value-driven effort to build a massive multilingual corpus	Open, transparent, and reproducible dataset assembled from diverse sources
Quality appraisal	Used web pages linked from Reddit with at least 3 karma as heuristic for quality	Filtered Common Crawl data with an off-the-shelf tool (JusText); other datasets used without filtering	Used rule-based filters intended to remove repetitive content, SEO pages, page code, etc.	Used rule-based quality filters, including heuristics from past work such as Gopher and C4
Toxic language appraisal	Filtered using a block list of "sexually explicit and otherwise offensive content"	Mostly documented rather than filtered toxic language; noted some intentional exclusions of potential sources	Filtered out documents with a high ratio of flagged terms for different languages	Customized toxicity filters per corpus, including classifiers and rule-based filters
Privacy appraisal	Not addressed in public documentation	Not specifically addressed but emphasized that all data were publicly available	Deduplicated and removed emails, social media handles, and IP addresses	Deduplicated and removed email addresses, phone numbers, IP addresses; invites removal requests
Data contamination appraisal	Removed Wikipedia articles due to presence in evaluation data	Noted a concern about data contamination but chose not to address	Not addressed in public documentation	Removed training documents with paragraphs present in select evaluation data
Documentation and finding aids	Published a list of the top domains and frequency in dataset but did not provide comprehensive public documentation	Individual subcorpora documented in some detail; released code for obtaining or replicating some of them	Documented many details of appraisal decisions in paper; created and released ROOTS search tool for interactive exploration	Documented many details of appraisal decisions; released replication code; included in WIMBD corpus exploration tool
Current status and downstream impacts	Used to train GPT2; Web Text neverreleased but was separately replicated as OpenWebText, which was used to train RoBERTa and other models, and was included in The Pile[12]	Used for training many LLMs, including LLaMa and the authors' own GPT-Neo; eventually removed from web following a DMCA takedown notification due to the inclusion of Books3 but still widely cited and used	Used by authors to train the BLOOM model; large parts downloadable via Hugging Face; full corpus available upon request; BigScience community remains active in this space	Used to train OLMo model; dataset downloadable via HuggingFace with some authentication required; plan to update dataset over time in response to personal information removal requests

While pretraining datasets exist along other forms of web archives, the inclusion of text in a pretraining dataset increases the odds that it will be preserved.[45,46] For example, the popular C4 dataset is based on the April 2019 Common Crawl snapshot,[47] which has therefore been duplicated many times, meaning that these particular data are unlikely to be lost over time.

Importantly, however, most pretraining datasets are extremely limited representations of original sources (e.g., containing text but not images, etc.), meaning that they are in no way an appropriate substitute for more traditional web archives. In addition, the characterization of datasets as "general purpose" (e.g., Gao et al.[12]) frames how people will encounter and use them and may increase the symbolic authoritative power of these datasets. Although this framing may imply that these datasets are comprehensive, data are intentionally and incidentally excluded during appraisal, as will be discussed in the following section.

Finally, dataset creation can also have powerful effects on future practices. Model developers commonly copy and build upon past approaches,[1,48] and datasets can be hard to meaningfully retract once they have been disseminated.[36,49,50] For example, although academic researchers created the BookCorpus dataset for training a multimodal sentence similarity model,[51] researchers at Google later reused it to augment Wikipedia as unlabeled pretraining data for their popular LLM, BERT.[52] BookCorpus was reused for several LLMs building on BERT[53,54] but was subsequently criticized for including problematic content and likely violating copyright restrictions.[53] Although the original authors no longer host or distribute the BookCorpus data, versions of it are still available from other sources. Similarly, Common Crawl has become a crucial resource for dataset builders, though approaches to filtering it vary, as we will discuss in the next section, particularly regarding toxic language. Importantly, and in contrast to traditional archives, most pretraining datasets include few or no metadata about context or provenance and do little to help users navigate them.

Given the impacts of pretraining data on knowledge production, the careful attention of researchers is needed on both the management of pretraining data and the practices that shape its development. In our discussion, we identify directions for future research that support the study and management of pretraining data using archival perspectives. First, however, we will look more closely at the practices that shape the development of these datasets.

APPRAISAL IN MAINSTREAM APPROACHES TO PRETRAINING DATA PROBLEMS

Even though they are central to the creation and evaluation of LLMs, pretraining dataset creators have not prioritized assembling pretraining datasets with the same level of care and detail as is done for traditional archives.[55,56] Nevertheless, there is a close parallel between an archivist's act of appraisal (assessing the value of a document in terms of it being worthy of preservation) and the practices of those who build these datasets. Appraisal is considered a central function of archival work, as it guides all other decisions about selection, preservation, and availability.[57,58]

Appraisal criteria are contextually dependent on how the archive is intended to be used. Those who build pretraining datasets make choices and evaluate (i.e., appraise) data with the primary goal of improving model performance on downstream tasks. Data are also often appraised and selected for inclusion with respect to a few key features, such as privacy vulnerabilities and toxic language.[1,13,59,60] Because of the scale involved, much of this appraisal is done via algorithmic filtering. Though archival studies emphasize the importance of documenting the principles and choices involved in appraisal and selection,[57,61] most pretraining datasets provide relatively little information about how or why these choices were made.[1] While there are exceptions—the creators of The Pile, for example, explain their position on copyright and fair use, as well as providing reasons for some exclusions[12]—this is not the norm.

Nevertheless, the community has converged on a few key issues for appraisal. For example, it has become common to evaluate the quality of text data when making selection decisions for inclusion in a pretraining corpus.[10,62–64] We can think of this as a measurement of a latent property of the text (i.e., "high quality" vs. "low quality"); however, the notion of quality is ambiguous and often unspecified. One popular way of operationalizing it is in terms of similarity to text that has been deemed high quality by other LLM researchers, such as Wikipedia.[1,10,64,65] This sort of approach entails an often-implicit yet specific set of value judgments. In particular, quality filters of this sort have been shown to systematically select against text written by certain groups of people,[54,66] including those from "poorer, less educated, rural areas,"[54] thereby reflecting implicit decisions about whose language should be included. In the rest of this section, we explore three additional examples of problems, like data quality, that pretraining dataset creators commonly face in appraising data and discuss parallels with issues in archival studies. Although these problems are often approached as being purely technical within research on LLMs, we emphasize that they are also inherently value-laden. To show this, we draw on theories of measurement and validity[67,68] to unpack how the problems being addressed are formulated and the value-laden assumptions carried by these formulations. In doing so, we show the limitations of these approaches and how more careful attention is needed from researchers who study the social impacts of technology on pretraining data appraisal. These kinds of appraisal decisions are important: as we will revisit in the discussion, researchers must also turn their attention to the effects of these decisions beyond model performance and behavior.

Toxic language

Many pretraining dataset creators evaluate (i.e., appraise) text for inclusion (i.e., selection) according to some standard of toxicity (e.g., Gao et al.,[12] Henderson et al.,[59] Penedo et al.[62]). Measuring, or detecting, toxic language is a key example of how algorithmic appraisal is done in practice. Following common practice among pretraining dataset creators, within this section, we do not distinguish between toxic language, hate speech, offensive language, and related topics, although these are distinct topics.[2,69] Beyond LLMs, toxic language detection has become a canonical task in natural language processing (NLP) (e.g., Waseem and Hovy,[70] Fortuna and Nunes,[71] Fortuna et al.,[72] Liang et al.[2]). However, measurement tasks such as appraising toxicity necessarily require making assumptions.[67,73] That is, underlying this work are questions about what should

count as toxic and according to whom: extensive research has identified many limitations of this task that speak to its specific set of assumptions, including the assumptions that toxic language is measurable outside of context[74,75] and that it is equivalently identifiable by everyone.[69,76] Because these assumptions imbue values in datasets and models, they deserve considerable attention from researchers.

For appraisal and selection of pretraining data, dataset creators have used a variety of approaches to assess toxicity (e.g., Brown et al.,[10] Henderson et al.,[59] Penedo et al.[62]). However, researchers have found that these methods may exhibit bias against language by and about marginalized social groups[77] and may be limited in their effectiveness.[4,77,78] For example, in creating C4 from Common Crawl, the creators excluded any document that contained any word on a list of "bad words."[47] However, a later investigation found that not only was this method ineffective at removing harmful language, it also disproportionately excluded text mentioning sexual minorities, as well as African American English and Hispanic-aligned English in comparison to White-aligned English.[77]

Archivists confront similar issues when dealing with toxic and offensive material in existing archives that were established hundreds of years ago. For example, archivists contend with the impact of hateful language in colonial archives on members of formerly colonized groups and grapple with the impact on historical narratives written by historians who use these archives. To manage these stakes, archivists intervene by consulting with experts and members of impacted groups to identify, mark, and sometimes annotate offensive language in archives, in addition to developing documentation that contextualizes these collections.[79–83] Additionally, archivists facilitate and advocate for community archives, partially to serve as counterevidence to harmful content in existing archives.[84,85] In community archives, archivists use participatory methods to determine appraisal criteria, recognizing that the identification of toxic or harmful materials and the decision to include these materials is both subjective and socially contextual.[84,86,87]

These issues speak to the fact that appraising data on the basis of its toxicity is necessarily a value-laden process: choosing to exclude data from a corpus based on the presence of words on a list is at least partly a decision about who or what matters. On the other hand, allowing unrestricted language into a pretraining dataset and releasing it as such has the potential to elevate such content and promote its circulation and reuse in knowledge production. Thus, more careful attention is needed on this process, including from scholars outside of natural language processing and machine learning. At a minimum, an archival perspective emphasizes the need for better documentation of both how and why exclusions were made and ideally incorporates more nuanced context into such decisions. Moreover, given the subjective nature of toxic language, it is important to underscore the power that researchers and engineers have in making these decisions.

Privacy vulnerabilities

Researchers who study and build pretraining datasets have also sought to mitigate privacy vulnerabilities (see Table 2 for examples).[59,60] For example, Carlini et al.[5] found that GPT-2 can generate personally identifiable information (PII) from its pretraining data, including names, phone numbers, and email addresses, and that frequently duplicated sequences are at higher risk of being generated. While LLM research has also looked to other solutions like memorization filters, it has become common to mitigate privacy risks via removal or redaction. As a result, many pretraining dataset creators attempt to filter out PII (operationalizing privacy risk as the presence of that information) as well as to deduplicate data (operationalizing privacy risks as the expected reproduction of certain data). As with identifying and removing toxic language, identifying privacy risks and mitigating them in this way are important acts of appraisal and selection.

Attention to the technical nuances of deduplication has revealed potential trade-offs. For instance, research shows that deduplicating at the sequence, rather than the document, level protects against some attacks,[90] while others have shown that more robust deduplication methods are needed.[91] Moreover, recent work has shown that deduplicating pretraining data can increase models' vulnerability to side-channel attacks but remains an important mitigation against common privacy vulnerabilities.[92]

More fundamentally, the question of what counts as PII or duplication—and whether these are even sufficient to address privacy concerns—is often unaddressed in this work. Measuring privacy vulnerabilities with PII and duplicates assumes that privacy is discrete and that privacy leakages are the only form of privacy risk. These assumptions are challenged by scholars who argue that privacy violations are contextual[93] and therefore find these approaches to appraising data insufficient.[94] Any operationalization of privacy vulnerabilities or their mitigation (PII, duplication, document removal) entails unstated assumptions about individuals, harms, and the costs of in/exclusion.

Scholars in archival studies (as well as library and information sciences broadly) contend seriously with the contextual nature of privacy and debate best practices for appraising publicly available personal data. Rather than focusing singularly on mitigating a narrow form of downstream privacy leakages, scholars advocate for appraising publicly available data with consideration of data subjects' perspectives.[95–97] Also emphasized is the need for ethical deliberation between data collectors and review boards[97–99] when appraising publicly available data, recognizing the value-laden and subjective nature of this task.

Evaluation and data contamination

In addition to imbuing values in pretraining data, the lack of careful attention and documentation in current appraisal practices makes it difficult to know what exactly is in these datasets. This complicates LLM evaluation, as it is difficult to know whether evaluation data might also be present in pretraining data. Contamination between evaluation data and pretraining data has implications both for rigorous evaluation and for using models for sociocultural analysis.[33] Rigorous LLM evaluation is critical, as evaluation drives the development of these models and is the basis for claims about LLM capabilities, which inform the way these models are used. As such, issues of evaluation illustrate the need for further study and documentation of pretraining data appraisal.

Although many researchers have found evaluation data in pretraining datasets[77] and agree that data contamination is a

problem for evaluation, the community has not yet established agreed-upon best practices for dealing with it. A common approach to assessing whether there is overlap between pretraining and test data is to use simple string matching (e.g., GPT-4 Technical Report[9]). More sophisticated techniques have been proposed, but all rely on strong assumptions.[100,101] Moreover, duplicates are themselves a complex construct,[102] and close but inexact matches might still have significant impacts on evaluation.[68,91,103–106] Thus, more research on data contamination is needed to support meaningful model evaluation.

Data contamination is sometimes considered in appraisal, with researchers using filters to exclude evaluation data from pretraining datasets (see Table 2). However, as new tasks, evaluation datasets, and methods are constantly being developed and used, this approach is insufficient to ensure that evaluation is not impacted by data contamination. Thus, researchers who evaluate LLMs need greater transparency of pretraining datasets (e.g., data and appraisal documentation) to support the validity of their evaluations. Although there is not an especially close parallel to the evaluation problem in traditional archives, archivists are nevertheless used to dealing with collections that are too large to be properly documented or organized. Work on developing tools for navigating collections, such as finding aids for archives, may be relevant here.[21,26] Archivists also take seriously the impact of their decisions on future research by others, as material that is not selected for inclusion in a traditional archive may end up being destroyed, potentially limiting what can be known in the future.[15,20] For the problem of pretraining datasets and LLM evaluation, new approaches are needed for exploring and navigating pretraining datasets, measuring duplication, and even rethinking how best to evaluate models.

DISCUSSION

While pretraining datasets are not archives in the traditional sense, the study of archives provides a useful theoretical framework for understanding the power of pretraining datasets in knowledge production. We hope to stimulate discussion among dataset creators and inspire research that will help us better understand and confront this power. As it stands, dataset creators are guided by relatively little empirical evidence when appraising pretraining data,[1] and the question of how access to these datasets should be managed is not straightforward. This work would be better supported by the research and perspectives of scholars who study the social impacts of technology from archival studies; information, computer, and social sciences; and related disciplines. Here, using lessons from archival studies, we identify opportunities for research that will support more responsible practices for creating and using pretraining datasets, thereby serving as a bridge for relevant communities of researchers.

As has been previously observed, the machine learning community broadly tends to use a "laissez-faire" approach to data collection with little regard for archival principles, transparency, or ethics.[7] As a snapshot of existing approaches, we summarize how the above concerns are (and are not) addressed in four prominent datasets in Table 2. By considering the practices archivists have developed to manage the power of the archive and why these practices were developed and adopted, several directions suggest themselves as particularly important. Key among these are documentation, transparency, and participation.[8,57,58]

Motivating critical inquiry on pretraining datasets

For context, it is worth considering what sparked and sustains archival studies' interrogation of the power wielded by archivists and archives. As historians turned to the archives looking to write the histories of marginalized social groups or about the everyday lives of people, archivists recognized the absence of relevant materials and began to contend with the impact of the "silence of the archives" on the historical record.[24,25,107] As a result, archival scholars have critically interrogated their practices and developed methods for responsibly managing the power they wield over decades.[15,29,30]

Evidence of the direct impacts of pretraining data curation, including carefully constructed model evaluations and audits of filtering algorithms, may motivate more careful attention to this process and its outputs. Researchers should study the impacts of pretraining data at the site of its reuse and in the context of model deployment to further our understanding of these datasets' impacts. For example, several studies reviewed in the preceding section found that algorithmic filtering techniques systematically exclude language by and about marginalized social groups. Since this silencing may not be easily mitigated through downstream interventions, evidence of its impacts may motivate more careful attention to pretraining data. To support this work, more research on where and how pretraining data and models are used and deployed is needed.

Documentation

Documentation should include not just what data were used for pretraining (including dates or version numbers, as appropriate) but also why and how data were chosen, appraised, and excluded. As archival scholars note, documenting this contextual information helps those who use a collection understand its meaning[108] and may impact the way a dataset is reused.[109] Even when the data itself cannot be shared, documenting and reporting on such decisions can be valuable, for example in helping those who use LLMs understand the limitations of these models (see, e.g., Davidson and Freire,[110] Gebru et al.,[111] and Hills et al.[112]). Further research is needed to ensure that documentation meets the needs of dataset and model users[113] and can be inspired by similar research in archival studies.[114] Analysis and documentation of appraisal and selection practices should extend not just to the creators of corpora but also to key data providers: organizations like Common Crawl play a unique and powerful role in the LLM ecosystem[115] and thus can offer a key point of intervention. Explicit attention to appraisal processes encourages both better analysis and more diverse approaches.

Finding aids and transparency

There is also a need for better tools for navigating, querying, and assessing pretraining corpora. Here archivists' work designing interfaces to make archives more user friendly and creating finding aids to provide users with information about a collections' materials, source, and structure can stand as inspiration.[21,27] For

pretraining data, tools like WIMDB,[60] which helps researchers measure and compare aspects of pretraining datasets (e.g., duplicate and synthetic text, PII, toxic language, and evaluation data contamination), and the ROOTS Search Tool,[31] which allows users to directly search the ROOTS pretraining dataset,[89] are excellent examples. However, further research is needed to make these tools comprehensive and properly matched to users' needs.

Developing methods for easily locating individual parts of a pretraining corpus in their original contexts may prove illuminating for many researchers, including those who study data contamination and model evaluation.[31] These tools may also help bring more perspectives on dataset creation to the surface. By revealing the contents of pretraining datasets to a wider audience, these tools may spark public conversation about whose data are and are not included (e.g., sensitive data, language dialects). In turn, this discourse may inform dataset creation practices or motivate participatory appraisal practices.

Participation and appraisal

In general, work in library and archival studies on community-driven and participatory archives[107,116,117] may be useful for developing appraisal practices for pretraining datasets. Past work has proposed a greater emphasis on community-contributed pretraining datasets,[7] which were used to some extent in building the ROOTS corpus,[89] but more of this work is needed. However, without critical reflection, scholars have made clear that such approaches can and will only add to the epistemic burden placed on already marginalized communities in data work.[118,119] While full-fledged community-driven datasets may seem impractical at the scale of pretraining data, novel technical approaches, such as modular and decentralized models[120] that make use of smaller-scale pretraining datasets, may make community-driven datasets viable and deserve additional consideration by researchers and practitioners who build LLMs.

Furthermore, other approaches that elicit community input to develop appraisal criteria may be feasible for large-scale pretraining datasets. Archivists take similar approaches to appraising materials with community input when constrained by scale or cost,[84,86,87] as we discuss in our exploration of the appraisal of toxic language. For pretraining datasets, researchers could use participatory methods to collectively identify criteria for appraising data on the basis of its toxicity or quality. These criteria could be used to develop scalable measures of toxicity or quality that incorporate more perspectives and nuanced forms of context.

This kind of community engagement is relevant to privacy appraisal as well. The collection of sensitive data for pretraining corpora, whether or not these data leak, may violate privacy expectations and laws.[40,93] Limiting pretraining data to what are publicly available might seem like a straightforward way of respecting privacy, but determining what should count as public (or was intended to be public) may still be challenging.[96]

Data could instead be appraised with consideration of data subjects' perspectives, following calls of archival and information scholars, as considered in our exploration of privacy appraisal practices. In the context of pretraining data, more research is needed on data subjects' perspectives to support this kind of appraisal. As research suggests that data subjects' perspectives on data collection shift depending on the sensitivity of the data and the context of its use, scholars advocate for research communities (i.e., LLM developers) to inform data subjects' perspectives by collectively engaging in public education efforts on the uses and risks of data.[95–97]

An archival perspective demands that we recognize and confront the power dynamics at play in the creation and circulation of pretraining resources. We hope that by recognizing the degree to which consequential choices are being made by a small set of people, scholars from a variety of disciplines will be encouraged to pursue research on this topic.

CONCLUSION

The selection of pretraining data for LLMs has been largely attended to as an engineering exercise. Yet, the curation of pretraining data is also a political process where both the artifact itself (the pretraining data) and any models trained on it will have cultural and political impacts that tend not to be widely considered. We adopt an archival perspective on pretraining data suggesting consideration of their power as informal archives and the processes that generate them. We highlight how common practices for LLMs have organized around addressing particular problems, such as mitigating specific privacy harms, which turn out to be practices of appraisal. The archival perspective points to the sources of power in the engineering choices in and around pretraining data. Ultimately, this framework offers a path forward to study not just the data but the systems that produce them.

ACKNOWLEDGMENTS

This paper originated as a workshop paper (https://openreview.net/forum?id=9xhUufywBX) presented at the Socially Responsible Language Modelling Research Workshop at the 2023 Conference on Neural Information Processing Systems. We are grateful to the *Patterns* editorial team for the opportunity to further develop this paper. We would like to thank Maria Antoniak, Kevyn Collins-Thompson, Jeremy Seeman, Amina Abdu, John Rudnik, and anonymous reviewers for their helpful feedback and suggestions.

AUTHOR CONTRIBUTIONS

All authors contributed intellectually to the development of ideas, analysis, and drafting the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., et al. (2023). A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.13169.

2. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research. Featured Certification, Expert Certification. https://doi.org/10.48550/arXiv.2211.09110.

3. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of Risks posed by Language Models. FAccT '22. Seoul, Republic of Korea. In Proceedings of the 2022 ACM Conference on Fairness,

Accountability, and Transparency (Association for Computing Machinery), pp. 214–229. https://doi.org/10.1145/3531146.3533088.

4. Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N.A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2020 (Association for Computational Linguistics), pp. 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301.

5. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T.B., Song, D., Erlingsson, U., et al. (2021). Extracting Training Data from Large Language Models. In Proceedings of the 30th USENIX Security Symposium, *6* (USENIX Association), pp. 2633–2650.

6. Feng, S., Park, C.Y., Liu, Y., and Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Lin- guistics, *vol 1: Long Papers* (Association for Computational Linguistics), pp. 11737–11762. https://doi.org/10.18653/v1/2023.acl-long.656.

7. Jo, E.S., and Gebru, T. (2020). Lessons from archives: strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery), pp. 306–316. https://doi.org/10.1145/3351095.3372829.

8. Schoenebeck, S., and Conway, P. (2020). Data and Power: Archival Appraisal Theory as a Frame- work for Data Preservation. Proc. ACM Hum. Comput. Interact. *4*, 1–18. (CSCW2). https://doi.org/10.1145/3415233.

9. OpenAI (2022). GPT-4 Technical Report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.08774.

10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901.

11. Gokaslan, A., and Cohen, V. (2019). OpenWebText Corpus. https://skylion007.github.io/OpenWebTextCorpus/.

12. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The Pile: An 800Gb dataset of diverse text for language modeling. Preprint at arXiv. https://doi.org/10.48550/arXiv.2101.00027.

13. Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. (2024). Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.00159.

14. Cook, T. (1997). What is Past is Prologue: A History of Archival Ideas Since 1898, and the Future Paradigm Shift. Archivaria *43*, 17–63.

15. Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. Arch. Sci. (Dordr). *13*, 95–120. https://doi.org/10.1007/s10502-012-9180-7.

16. Bailey, J. (2013). Disrespect des fonds: Rethinking arrangement and description in born-digital archives. Archive Journal *3*, 201–212.

17. Schellenberg, T.R., et al. (1956). Modern Archives (University of Chicago Press Chicago).

18. Buckland, M.K. (1997). What is a "document"? J. Am. Soc. Inf. Sci. *48*, 804–809. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<804::AID-ASI5>3.0.CO;2-V.

19. Richards, T. (1993). The Imperial Archive: Knowledge and the Fantasy of Empire. In The Imperial Archive: Knowledge and the Fantasy of Empire (Verso Books).

20. Schwartz, J.M., and Cook, T. (2002). Archives, records, and power: The making of modern memory. Arch. Sci. *2*, 1–19. https://doi.org/10.1007/BF02435628.

21. Hedstrom, M. (2002). Archives, memory, and interfaces with the past. Arch. Sci. *2*, 21–43. https://doi.org/10.1023/A:1020800828257.

22. O'toole, J.M. (2002). Cortes's notary: The symbolic power of records. Arch. Sci. *2*, 45–61. https://doi.org/10.1007/BF02435630.

23. Jacobsen, T., Punzalan, R.L., and Hedstrom, M.L. (2013). Invoking "collective memory": Map- ping the emergence of a concept in archival science. Arch. Sci. (Dordr). *13*, 217–251. https://doi.org/10.1007/s10502-013-9199-4.

24. Trouillot, M. (1995). Silencing the Past: Power and the Production of History (Beacon Press books. Beacon Press).

25. Thomas, D., Fowler, S., and Johnson, V. (2017). The Silence of the Archive (Facet Publishing).

26. Yakel, E. (2003). Archival representation. Arch. Sci. *3*, 1–25. https://doi.org/10.1007/BF02438926.

27. Yakel, E. (2011). Who Represents the Past? In Archives, Records, and the Social Web". Controlling the Past: Documenting Society and Institutions, T. Cook, ed. (Society of American Archivists), pp. 258–278.

28. Ham, F. (1981). Archival strategies for the post-custodial era. Am. Arch. *44*, 207–216. https://doi.org/10.17723/aarc.44.3.6228121p01m8k376.

29. Caswell, M., and Cifor, M. (2016). From human rights to feminist ethics: radical empathy in the archives. Archivaria *81*, 23–43.

30. Punzalan, R.L., and Caswell, M. (2016). Critical directions for archival approaches to social justice. The Library Quarterly *86*, 25–42. https://doi.org/10.1086/684145.

31. Piktus, A., Akiki, C., Villegas, P., Laurençon, H., Dupont, G., Luccioni, S., Jernite, Y., and Rogers, A. (2023). The ROOTS Search Tool: Data Transparency for LLMs. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, *vol 3* (Association for Computational Linguistics), pp. 304–314. https://doi.org/10.18653/v1/2023.acl-demo.29.

32. Spennemann, D.H.R. (2023). ChatGPT and the Generation of Digitally Born "Knowledge": How Does a Generative AI Language Model Interpret Cultural Heritage Values? Knowledge *3*, 480–512. https://doi.org/10.3390/knowledge3030032.

33. Chang, K., Cramer, M., Soni, S., and Bamman, D. (2023). Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 7312–7327. https://doi.org/10.18653/v1/2023.emnlp-main.453.

34. Garcia1, G.G., and Weilbach, C. (2023). If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.10808.

35. 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*. https://doi.org/10.1177/20539517211035955.2.

36. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., and Hanna, A. (2021). Data and its (Dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research. Patterns *2*. https://doi.org/10.1016/j.patter.2021.100336.

37. Scheuerman, M.K., Weathington, K., Mugunthan, T., Denton, E., and Fiesler, C. (2023). From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. Proc. ACM Hum. Comput. Interact. *7*, 1–33. https://doi.org/10.1145/3579488.

38. De Vynck, G. (2023). ChatGPT Maker OpenAI Faces a Lawsuit over How it Used People's Data (The Washington Post). https://www.washingtonpost.com/technology/2023/06/28/openai-chatgpt-lawsuit-class-action/.

39. Lee, K., Cooper, A.F., and Grimmelmann, J. (2024). Talkin 'bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version). In Proceedings of the Symposium on Computer Science and Law. (Association for Computing Machinery). https://doi.org/10.1145/3614407.3643696.

40. Roberston, A. (2023). ChatGPT returns to Italy after ban (The Verge). https://www.theverge.com/2023/4/28/23702883/chatgpt-italy-ban-lifted-gpdp-data-protection age-verification.

41. Samuelson, P. (2023). Generative AI meets copyright. Science *381*, 158–161.

42. Small, Z. (2023). Sarah Silverman Sues OpenAI and Meta over Copyright Infringement (The New York Times). https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html.

43. Lepore, J. (2015). The Cobweb (the New Yorker). Accessed 2023-12-04. https://www.newyorker.com/magazine/2015/01/26/cobweb.

44. Bruns, A., and Weller, K. (2016). Twitter as a first draft of the present: And the challenges of preserving it for the future. In Proceedings of the 8th ACM Conference on Web Science (Association for Computing Machinery), pp. 183–189. https://doi.org/10.1145/2908131.2908174.

45. Milligan, I. (2016). Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. In Ternational Journal of Humanities and Arts Computing, *10*, pp. 78–94. https://doi.org/10.3366/ijhac.2016.0161.

46. Murphy, B. (2022). We the Dead: Preserving Data at the End of the World (University of North Carolina Press).

47. 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. J. Mach. Learn. Res. *21*, 5485–5551.

48. Zhou, K., Blodgett, S.L., Trischler, A., Daume', H., III, Suleman, K., and Olteanu, A. (2022). Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Lin- guistics), pp. 314–324. https://doi.org/10.18653/v1/2022.naacl-main.24.

49. Peng, K.L., Mathur, A., and Narayanan, A. (2021). Mitigating dataset harms requires steward- ship: Lessons from 1000 papers. In Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track.

50. Luccioni, A.S., Corry, F., Sridharan, H., Ananny, M., Schultz, J., and Crawford, K. (2022). A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Commu- nication. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Trans- parency (Association for Computing Machinery)), pp. 199–212. https://doi.org/10.1145/3531146.3533086.

51. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pp. 19–27. https://doi.org/10.1109/ICCV.2015.11.

52. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidi- rectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, *Volume 1* (Association for Computational Linguistics), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

53. Bandy, J., and Vincent, N. (2021). Addressing "documentation debt" in machine learning: A retrospective datasheet for BookCorpus". In Thirty-fifth Conference on Neural Information Pro- cessing Systems, Datasets and Benchmarks Track.

54. Gururangan, S., Card, D., Dreier, S., Gade, E., Wang, L., Wang, Z., Zettlemoyer, L., and Smith, N.A. (2022a). Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2562–2580. https://doi.org/10.18653/v1/2022.emnlp-main.165.

55. Gero, K.I., Das, P., Dognin, P., Padhi, I., Sattigeri, P., and Varshney, K.R. (2023). The incentive gap in data work in the era of large models. Nat. Mach. Intell. *5*, 565–567. https://doi.org/10.1038/s42256-023-00673-x.

56. 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 In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–15. https://doi.org/10.1145/3411764.3445518.

57. Samuels, H. (1986). Who Controls the Past*49* (The American Archivist), pp. 109–124.

58. Cook, T. (2011). In "Documenting Society and Institutions: The Influence of Helen Willa Samuels". Controlling the Past: Documenting Society and Institutions, T. Cook, ed. (Society of American Archivists), pp. 1–30.

59. Henderson, P., Krass, M.S., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D., and Ho, D.E. (2022). Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open- Source Legal Dataset. In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

60. Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Pete Walsh, D.G., Soldaini, L., Singh, S., Hajishirzi, H., et al. (2023). What's In My Big Data?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.20707.

61. Cox, R.J. (1994). The Documentation Strategy and Archival Appraisal Principles: A Different Per- Spective. Archivaria *38*.

62. 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. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.01116.

63. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training Gopher. Preprint at arXiv. https://doi.org/10.48550/arXiv.2112.11446.

64. Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzma'n, F., Joulin, A., and Grave, E. (2020). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the Twelfth Language Resources and Evaluation Conference (European Language Resources Association), pp. 4003–4012.

65. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al. (2022). PaLM: Scaling language modeling with pathways. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.02311.

66. Lucy, L., Gururangan, S., Soldaini, L., Strubell, E., Bamman, D., Klein, L., and Dodge, J. (2024). AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.06408.

67. Jacobs, A.Z., and Wallach, H. (2021). Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery)), pp. 375–385. https://doi.org/10.1145/3442188.3445901.

68. Subramonian, A., Yuan, X., Daume', H., III, and Blodgett, S.L. (2023). It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In Findings of the Association for Computational Linguistics 2023 (Association for Computational Linguistics), pp. 3234–3279. https://doi.org/10.18653/v1/2023.findings-acl.202.

69. Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N.A. (2019). The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 1668–1678. https://doi.org/10.18653/v1/P19-1163.

70. Waseem, Z., and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop (Association for Computational Linguistics), pp. 88–93. https://doi.org/10.18653/v1/N16-2013.

71. Fortuna, P., and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Comput. Surv. *51*, 1–30. https://doi.org/10.1145/3232676.

72. Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In Proceedings of the Twelfth Language Resources and Evaluation Conference (European Language Resources Association), pp. 6786–6794.

73. Blodgett, S.L., Barocas, S., Daume', H., III, and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485.

74. Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., and Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 4296–4305. https://doi.org/10.18653/v1/2020.acl-main.396.

75. Aken, B. van, Risch, J., Krestel, R., and Löser, A. (2018). Challenges for Toxic Comment Clas- sification: An In-Depth Error Analysis. In Proceedings of the 2nd Workshop on Abusive Lan- guage Online (Association for Computational Linguistics), pp. 33–42. https://doi.org/10.18653/v1/W18-5105.

76. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N.A. (2022). Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In Pro- ceedings of the 2022 Conference of the North American Chapter of the Association for Compu- tational Linguistics: Human Language Technologies (Association for Computational Linguistics), pp. 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431.

77. Dodge, J., Sap, M., Marasovic', 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 (Association for Computational Linguistics), pp. 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98.

78. Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., and Huang, P.-S. (2021). Challenges in Detoxifying Language Models. In Find- ings of the Association for Computational Linguistics: EMNLP 2021 (Association for Compu- tational Linguistics), pp. 2447–2469. https://doi.org/10.18653/v1/2021.findings-emnlp.210.

79. Aboriginal and (ATSIDA), T. S. I. D. A. (2013). ATSIDA Protocols. https://www.atsida.edu.au/protocols/atsida.

80. The National Archives (2023). Cataloguing Physical Records: Guidance for Government Depart- Ments. https://cdn.nationalarchives.gov.uk/documents/information-management/cataloguing-physical-records-guidance-for-government-departments.pdf.

81. Chilcott, A. (2022). Towards Protocols for Describing Racially Offensive Language in UK Public Archives. In Archives in a Changing Climate-Part I & Part II (Springer), pp. 151–168. https://doi.org/10.1007/s10502-019-09314-y.

82. Underhill, K.J. (2006). Protocols for Native American archival materials. RBM A J. Rare Books, Manuscripts, Cult. Herit. (Chic.) *7*, 134–145. https://doi.org/10.5860/rbm.7.2.267.

83. Punzalan, R.L., Marsh, D.E., and Cools, K. (2017). Beyond Clicks, Likes, and Downloads: Iden- Tifying Meaningful Impacts for Digitized Ethnographic Archives. Archivaria *84*, 61–102.

84. Caswell, M. (2014). Toward a survivor-centered approach to records documenting human rights abuse: lessons from community archives. Arch. Sci. (Dordr). *14*, 307–322. https://doi.org/10.1007/s10502-014-9220-6.

85. Caswell, M., Migoni, A.A., Geraci, N., and Cifor, M. (2017). 'To be able to imagine otherwise': community archives and the importance of representation. Archives and Records *38*, 5–26. https://doi.org/10.1080/23257962.2016.1260445.

86. Caswell, M. (2014b). Inventing New Archival Imaginaries: Theoretical Foundations for Identity- Based Community Archives. In Identity Palimpsests: Archiving Ethnicity in the U.S. and Canada, D. Dominique and A. Levi, eds. (Litwin Books), pp. 35–53.

87. Zavala, J., Migoni, A.A., Caswell, M., Geraci, N., and Cifor, M. (2017). 'A process where we're all at the table': Community archives challenging dominant modes of archival practice. Arch. Manuscripts *45*, 202–215. https://doi.org/10.1080/01576895.2017.1377088.

88. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog *1*, 9. https://api.semanticscholar.org/CorpusID:160025533.

89. Laurenc,on, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., Von Werra, L., Mou, C., Gonza'lez Ponferrada, E., Nguyen, H., et al. (2022). The BigScience ROOTS Corpus: A 1.6TB composite multilingual dataset. Adv. Neural Inf. Process. Syst. *35*, 31809–31826.

90. Kandpal, N., Wallace, E., and Raffel, C. (2022). Deduplicating training data mitigates privacy risks in language models. In International Conference on Machine Learning (PMLR), pp. 10697–10707. https://doi.org/10.48550/arxiv.2202.06539.

91. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2022). Deduplicating Training Data Makes Language Models Better. In Proceedings of the 60th An- nual Meeting of the Association for Computational Linguistics, *Volume 1* (Association for Computational Linguistics), pp. 8424–8445. https://doi.org/10.18653/v1/2022.acl-long.577.

92. Debenedetti, E., Severi, G., Carlini, N., Choquette-Choo, C.A., Jagielski, M., Nasr, M., Wallace, E., and Trame'r, F. (2023). Privacy Side Channels in Machine Learning Systems. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.05610.

93. Nissenbaum, H. (2004). Privacy as Contextual Integrity. Wash. L. Rev. *79*, 119.

94. Brown, H., Lee, K., Mireshghallah, F., Shokri, R., and Trame'r, F. (2022). What does it mean for a language model to preserve privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery), pp. 2280–2292.

95. Hemphill, L., Schöpke-Gonzalez, A., and Panda, A. (2022). Comparative sensitivity of social media data and their acceptable use in research. Sci. Data *9*, 643. https://doi.org/10.1038/s41597-022-01773-w.

96. Fiesler, C., and Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. Soc. Media Soc. *4*. 205630511876336. https://doi.org/10.1177/2056305118763366.

97. Vitak, J., Shilton, K., and Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical chal- lenges, practices, and beliefs in the online data research community. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (Association for Computing Machinery), pp. 941–953. https://doi.org/10.1145/2818048.2820078.

98. Reardon, S., Samberg, R., and Vollmer, T. (2021). Building Legal Literacies for Text Data Min- Ing (University of California Berkeley Library). Chap. 6, Ethics. https://doi.org/10.48451/S1159P.

99. Heise, A.H.H., Hongladarom, S., Jobin, A., Kinder-Kurlanda, K., Sun, S., Lim, E.L., Markham, A., Reilly, P.J., Tiidenberg, K., and Wilhelm, C. (2019). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf.

100. Oren, Y., Meister, N., Chatterji, N., Ladhak, F., and Hashimoto, T.B. (2023). Proving Test Set Contamination in Black Box Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.17623.

101. Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. (2023). Detecting Pretraining Data from Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.16789.

102. Yauney, G., Reif, E., and Mimno, D. (2023). Data Similarity is Not Enough to Explain Language Model Performance. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 11295–11304. https://doi.org/10.18653/v1/2023.emnlp-main.695.

103. Raji, D., Denton, E., Bender, E.M., Hanna, A., and Paullada, A. (2021). AI and the Everything in the Whole Wide World Benchmark. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks.

104. Blevins, T., and Zettlemoyer, L. (2022). Language Contamination Helps Explains the Cross- lingual Capabilities of English Pretrained Models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 3563–3574. https://doi.org/10.18653/v1/2022.emnlp-main.233.

105. Razeghi, Y., Logan IV, R.L., Gardner, M., and Singh, S. (2022). Impact of pretraining term frequencies on few-shot numerical reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2022 (Association for Computational Linguistics)), pp. 840–854. https://doi.org/10.18653/v1/2022.findings-emnlp.59.

106. Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. (2024). Do Membership Inference Attacks Work on Large Language Models?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.07841.

107. Caswell, M., Cifor, M., and Ramirez, M.H. (2016). "To suddenly discover yourself existing": uncovering the impact of community archives. Am. Archivist *79*, 56–81. https://doi.org/10.17723/0360-9081.79.1.56.

108. Ketelaar, E. (2001). Tacit narratives: the meanings of archives. Arch. Sci. *1*, 131–141.

109. Pasquetto, I.V., Randles, B.M., and Borgman, C.L. (2017). On the reuse of scientific data. Data Sci. J. *16*, 1–9. https://doi.org/10.5334/dsj-2017-008.

110. Davidson, S.B., and Freire, J. (2008). Provenance and scientific workflows: challenges and op- portunities. In Proceedings of the 2008 ACM SIGMOD international conference on Manage- ment of data, pp. 1345–1350. https://doi.org/10.1145/1376616.1376772.

111. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., and Crawford, K. (2021). Datasheets for datasets. Commun. ACM *64*, 86–92. https://doi.org/10.1145/3458723.

112. Hills, D., Downs, R.R., Duerr, R., Goldstein, J.C., Parsons, M.A., and Ramapriyan, H.K. (2015). The Importance of Data Set Provenance for Science, Eos. Advancing Earth and Space Sciences *96*.

113. Heger, A.K., Marquis, L.B., Vorvoreanu, M., Wallach, H., and Wortman Vaughan, J. (2022). Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Chal- lenges, and Desiderata. Proc. ACM Hum. Comput. Interact. *6*, 1–29.

114. Faniel, I.M., Frank, R.D., and Yakel, E. (2019). Context from the data reuser's point of view. J. Doc. *75*, 1274–1297.

115. Baack, S., and Insights, M. (2024). Training Data for the Price of a Sandwich (Mozilla Foundation). Accessed 2023-2-08. https://assets.mofoprod.net/network/documents/Common_Crawl_Mozilla_Foundation_2024.pdf.

116. Flinn, A., Stevens, M., and Shepherd, E. (2009). Whose memories, whose archives? Independent community archives, autonomy and the mainstream. Arch. Sci. (Dordr). *9*, 71–86. https://doi.org/10.1007/s10502-009-9105-2.

117. Huvila, I. (2008). Participatory archive: Towards decentralised curation, radical user orientation, and broader contextualisation of records management. Arch. Sci. (Dordr). *8*, 15–36. https://doi.org/10.1007/s10502-008-9071-0.

118. Pierre, J., Crooks, R., Currie, M., Paris, B., and Pasquetto, I. (2021). Getting Ourselves Together: Data-centered participatory design research & epistemic burden. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–11. https://doi.org/10.1145/3411764.3445103.

119. Sloane, M., Moss, E., Awomolo, O., and Forlano, L. (2022). Participation is not a design fix for machine learning. In Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (Association for Computing Machinery), pp. 1–6. https://doi.org/10.1145/3551624.3555285.

120. Gururangan, S., Lewis, M., Holtzman, A., Smith, N.A., and Zettlemoyer, L. (2022b). DEMix Layers: Disentangling Domains for Modular Language Modeling. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Hu- man Language Technologies (Association for Computational Linguistics), pp. 5557–5576. https://doi.org/10.18653/v1/2022.naacl-main.407.