An Archival Perspective on Pretraining Data

Meera A. Desai
madesai@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Abigail Z. Jacobs
azjacobs@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Dallas Card
dalc@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Abstract

Research in NLP on pretraining data has largely focused on identifying and mitigating downstream risks in models. We argue that more critical attention is needed to pretraining datasets and the systems that produce them. To highlight the broader range of impacts of pretraining corpora, we consider the analogy between pretraining datasets and archives. Within the broader ecosystem of datasets and models, we focus especially on processes involved in the creation of pretraining data. By adopting an archives perspective, we surface impacts beyond directly shaping model behavior, including the role of pretraining data corpora as independent data artifacts and the ways that their collection shape future practices. In particular, we explore research in NLP that parallels archival practices of appraisal: we consider the practices of filtering of pretraining data and critically examine the problem formulations taken on by this work. In doing so, we underscore how choices about what is included in pretraining data are necessarily subjective decisions about values. We conclude by drawing on archival studies to offer insights on paths forward.

1 Introduction

Research in NLP has increasingly focused on downstream risks from large language models (LLMs), such as discrimination or the generation of toxic or misleading content [55, 54, 99]. Some research has considered how these risks relate to pretraining data [34, 8, 55, 26], while model developers emphasize downstream interventions to these problems (e.g., reinforcement learning with human feedback [106, 61]). Yet pretraining datasets are important not just for their influence on model behavior, but as unique sociocultural collections with lasting impacts. Given their unique role in our information ecosystem, more critical attention to pretraining datasets and their production is needed.

Given archival studies' long tradition of contending with the impacts of assembling collections of sociocultural materials, an archival perspective offers useful insights into understanding the inherent power of pretraining data, those who create it, and the practices that shape its development [48, 83]. In particular, we argue that attending to pretraining data is important, and requires studying not just the data, but also the systems that produce it.

In this paper we draw on the field of archival studies to discuss broader issues with the creation, representation, and circulation of pretraining datasets for LLMs. Although not archives in the traditional sense, we make use of the analogy to archives to theorize their importance and to interrogate how pretraining datasets are built and used. We first provide relevant background on archival studies (§2). We then consider the practices used to build pretraining datasets, focusing on three common concerns among LLM researchers. By drawing on the parallel between these practices and the archival practice of appraisal, we surface the values and assumptions underlying these areas of focus (§3). Finally, in §4, we discuss the implications of an archival perspective for the creation of pretraining datasets and their effects on the broader information ecosystem.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 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, 16]. 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 material was drawn [2, 16]. Facing ever larger collections, archivists developed the practices of *appraisal* and *selection* to evaluate and choose materials for preservation [81, 14].

Over time, the field of archival studies has come to recognize the power-laden nature of archives [16] and their authority to shape our knowledge and written history [75, 84, 41]. Since our understanding of the past mediates the formation of cultural memory and human identity [16, 63], 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 [94, 93]. In addition to appraisal and selection, archivists 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 [41, 102, 103]. Archival scholarship thus explores practices that embrace the inherent subjectivity of assembling an archive while attending to the political stakes at hand (e.g. [40, 56, 60]).

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., [36, 31, 44, 69]). An archival perspective suggests 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 [48, 83].

Pretraining data is most obviously relevant as training data for language models. The choices of data determines not only the capabilities of a generative model, but more fundamentally, the language affordances (e.g., English or multilingual) and the information it includes. To the extent that people use LLMs as interfaces into history and culture, the selection of data shapes and constrains that experience [89, 11, 32]. In parallel to the appraisal of material for archives, those who appraise and select information for archives are exercising an important form of power. 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 what is possible for associated models.

Most web-scale datasets are collected without consent, and datasets can be hard to meaningfully retract once they have been disseminated [64, 67, 82]. Automated selection is virtually guaranteed to include sensitive and contested material, and questions of copyright and ownership are actively being litigated [19, 51, 76, 78, 87]. Yet independent of downstream model use, the very act of including material in a pretraining corpus can bring attention to these materials and lend them legitimacy [21]. In addition, the common indication of datasets as "general purpose" (e.g., [31]) frames how people will encounter and use them. Moreover, such datasets will be replicated in pretraining (or other) datasets, making them likely to be more available in the future; what is not included is more likely to be lost over time [57, 58].

Dataset creation can also have powerful effects on future practices. Model developers commonly copy and build upon past work [55, 104]. For example, BookCorpus was created for training a sentence similarity model [105], but was later used to augment Wikipedia as unlabeled pretraining data for BERT [22]. BookCorpus was then reused by several LLMs building on BERT [3, 37], but was subsequently criticized for including problematic content and likely violating copyright restrictions [3]. Common Crawl has similarly become a crucial resource for dataset builders, though approaches to filtering it vary (see especially §3.1). Importantly, and in contrast to traditional archives, most pretraining datasets include little or no metadata about context or provenance, and do little to help users navigate them.

Given the impacts of pretraining data in the larger information ecosystem, more careful attention is needed to the both the management of pretraining data and the practices that shape its development. At the end of the paper, we will revisit the management of pretraining data using archival perspectives. First, however, we will look more closely at the practices that shape the development of these datasets.

3 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 archives. 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, in contrast, is considered a central function of archival work, as it guides all other decisions about selection, preservation, and availability [77, 15].

Those who build pretraining datasets make choices and evaluate data (i.e., appraisal) with the primary goal of improving model performance on desired downstream tasks. Data is also often appraised with respect to a few agreed upon problems, such as privacy vulnerabilities and toxic language [88, 55, 44, 25]. Because of the scale involved, much of this appraisal is done via algorithmic filtering. Though archival studies emphasizes the importance of documenting the principles and choices involved in appraisal and selection [17, 77], most pretraining datasets provide relatively little information about how or why these choices were made [55]. While there are exceptions—the creators of The Pile, for example, explain their position on copyright and fair use, and (for instance) note that they exclude the Congressional Record because it contains a high level of racist content [31]—this is not the norm.

Nevertheless, the community has converged on a few key issues for appraisal. For example, in selecting data to include, it has become common to evaluate "data quality" [6, 66, 70, 101]. We can think of this as a measurement of a latent property of the text with respect to some standard. 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", like Wikipedia [101, 6, 13, 55]. Without it being explicit, this sort of approach entails tradeoffs and value judgements. In particular, quality filters of this sort have been shown to systematically erase text written by people from "poorer, less educated, rural areas" [37].

In the rest of this section, we explore three additional examples of problems, like data quality, that pretraining dataset creators commonly use to appraise data via algorithmic filtering. Although these problems are often approached as purely technical, they are also inherently value-laden. To show this, we draw on theories of measurement and validity [47, 91] to unpack how the problems being addressed are formulated, and the value-laden assumptions carried by these formulation. In doing so, we show the limitations of these approaches, and how more careful attention is needed in the appraisal of pretraining data. These kinds of appraisal decisions are important: as we will revisit in the discussion (§4), we also need to think more broadly about the processes by which these datasets are created, and the effects they have beyond model training.

3.1 Toxic Language

Many pretraining dataset creators appraise (i.e., evaluate) text for inclusion according to some standard of toxicity[1] (e.g., [31, 44, 66]). As an evocative example of algorithmic appraisal, we show how this task (i.e., measuring "toxic language", and evaluating and then selecting data for inclusion) requires making assumptions, is imbued with value decisions, and ultimately requires more care and attention both from practitioners and the larger the research community .

Toxic language detection has become a canonical task in NLP beyond LLMs (e.g, [98, 29, 30, 54]). Underlying this work, however, 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 assumption that toxic language is measurable outside of context [65, 96], and that it is equivalently identifiable by everyone [79, 80].

A variety of approaches have been used to assess toxicity in pretraining data (e.g., [6, 44, 66]). However, researchers have critically examined these approaches, finding that these methods may exhibit bias against language by and about marginalized social groups [24] and may be limited in their effectiveness [24, 34, 100]. For example, in creating C4 from Common Crawl, the creators appraised (and selected) text by excluding any document that contained any word on a list of "bad words" [71]. However, a later investigation found that not only was this method ineffective at removing harmful language, it also tended to disproportionately excluded text mentioning sexual minorities, as well as African American English and Hispanic-aligned English in comparison to White-aligned English [24].

Archivists confront similar issues when dealing with toxic and offensive material in existing archives that were appraised 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

[1] Although they are distinct concepts, we do not distinguish within this section between toxic language, hate speech, offensive language, and related topics [54, 79].

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 [1, 92, 12, 95]. Additionally, archivists facilitate and advocate for community archives, partially to serve as counter evidence to harmful content in existing archives [9, 90, 16]. 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 are both subjective and socially contextual [9, 10].

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 the information ecosystem. As such, more careful attention is needed on this process. At a minimum, an archival perspective emphasizes the need for much greater documentation of both how and why exclusions were made, and ideally incorporating 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.

3.2 Privacy Vulnerabilities

Privacy vulnerabilities are another agreed upon problem among researchers who study and build pretraining datasets [44, 25]. For example, Carlini et al. (2021) find 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. As a result, it has become common practice to sanitize data of PII (operationalizing privacy risks as the presence of that information) as well as to deduplicate data (operationalizing privacy risks as the expected reproduction of certain data). While NLP research has also looked to other solutions like memorization filters, it has become common to mitigate privacy risks via such removal or redaction. Identifying privacy risks and mitigating them in this way are themselves acts of appraisal and selection.

Additional research has focused on the technical nuances of deduplication and its impacts. For instance, research shows that deduplicating at the sequence, rather than document, level protects against some attacks [49], while others have shown that more robust deduplication methods are needed [52]. On the other hand, more recent work has shown that deduplicating pretraining data can increase models' vulnerability to side-channel attacks, yet remains an important mitigation against common privacy vulnerabilities [20].

Although mitigating privacy vulnerabilities clearly involves technical challenges, 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. In particular, measuring privacy vulnerabilities with PII and duplicates assumes that privacy is discrete and that privacy leakages are the only form of privacy risk. The choice to operationalize privacy vulnerabilities in these terms (PII, duplication, document removal) entails unstated assumptions about individuals, harms, and the costs of in/exclusion. These assumptions are challenged by scholars who argue that privacy violations are contextual [59] and therefore find these approaches to appraising data insufficient [5].

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 [43, 27, 97]. Also emphasized is the need for ethical deliberation between data collectors and review boards [97, 74, 42] when appraising publicly available data, recognizing the value-laden and subjective nature of this task.

3.3 Evaluation and Data Contamination

Measuring duplication is also relevant to model evaluation, where data contamination has become a significant problem. In addition to generic issues like accounting for hyperparameter tuning [23], statistical power [7], and establishing meaningful evaluation metrics [104], evaluating LLMs entails additional challenges with ensuring clean separation between training and test data; because of how

pretraining datasets are created, it is difficult to know whether evaluation data might also be present in pretraining data. This has implications both for rigorous evaluation, and for using models for sociocultural analysis [11]. Although this is not something that is typically addressed at the stage of building pretraining corpora, it speaks to how data collection practices have destablized what had become standard evaluation practices.

Although researchers have agreed 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., [61]). More sophisticated techniques have been proposed, but all rely on strong assumptions [62, 85]. Moreover, as discussed in §3.2, duplicates are themselves a complex construct, and close but inexact matches might be still relevant for meaningful evaluation [52, 72, 4, 73, 91].

This issue speaks to the need for more careful attention to pretraining data: while better appraisal techniques can be developed for assessing overlap, part of the reason for this problem is the lack of understanding of what exactly is in these datasets, and the lack of tools to navigate them. Going forward, new approaches are needed for exploring pretraining datasets, measuring duplication, and rethinking evaluation data in relation to pretraining corpora.

4 Discussion

While pretraining datasets are not archives in the traditional sense, the study of archives provides a useful theoretical framework for understanding the power-laden nature of pretraining datasets. Research on the social impact of pretraining data has largely focused on mitigating risks through computational filtering techniques, which we argue are practices parallel to the archival practice of appraisal. We analyze these practices through a measurement lens, revealing how common appraisal problems such as toxic language, privacy, and data contamination, are operationalized using specific assumptions that enact narrow meanings of these complex constructs. The creation and use of pretraining data has impacts that require broader and more creative interventions to enact more substantive, valid appraisal.

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 every day 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 [93]. As a result, archival scholars have critically interrogated their practices and developed methods for responsibly managing the power they wield over decades [16].

Evidence of the direct impacts of pretraining data curation 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 how these datasets impact our world. For example, several studies reviewed in §3 found that algorithmic filtering techniques systematically exclude language by and about marginalized social groups. Since the impact of these silences in pretraining data may be more difficult to fix with downstream interventions than other harms, evidence of their impacts may motivate more careful attention to pretraining data. To support this work, more research on where and how pretraining data and models are (re)used and deployed is needed.

As has previously been 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 [48]. By considering the practices archivists have developed to manage the power of the archive, several directions suggest themselves as particularly important. Key among these are documentation, transparency, privacy appraisal, and participation [77, 15, 83].

Documentation should include not just what data was used for pretraining (including dates or version numbers, as appropriate), but also what assumptions and values were used for choosing, appraising, and excluding data. There is value in documenting and reporting on such decisions, even when the data itself cannot be shared (see, e.g., [18, 33, 45]). 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, and thus can offer a key point of intervention. Explicit attention to appraisal processes encourages both better analysis and more diverse approaches.

There is also a need for better tools for navigating, querying, and assessing pretraining corpora. This echoes archivists' work on creating finding aids and interfaces, so their work can also stand as inspiration [41, 103]. For pretraining data, tools like WIMDB [25] and the ROOTS Search Tool [50, 69] are excellent examples, but more work is needed to make these tools comprehensive and properly matched to users' needs. More generally, developing ways to easily locate individual parts of a pretraining corpus in their original context might prove illuminating, as well as bolster documentation, openness, and transparency efforts. For example, with a well designed finding aid, for pretraining data owners could invite participatory documentation, drawing on similar efforts in archival studies [39, 12].

At the same time, archivists have long recognized that there are no simple answers to what data should be collected, preserved, and made available to the public. The collection of sensitive data for pretraining corpora, whether or not this data leaks, may violate privacy expectations and laws [76, 59]. Limiting pretraining data to what is publicly available might seem like a straightforward way of respecting privacy, but determining what should count as public (or intended to be public) may still be challenging [27].

As an alternative, data could instead be appraised with consideration of data subjects' perspectives, following calls of archival and information scholars. 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 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 [27, 43, 97]. Appraisal based on other frameworks, such as legal principles [44], can also offer complementary ways to approach these problems.

Others in library and archival studies have worked extensively on community-driven and participatory archives [28, 46]. Scholars have argued that especially in contexts with historical power imbalances between people, such as decolonial contexts, people whose documents are being stored and made available should have some input into how this material is selected, treated, presented, and made available [35, 56]. Past work has proposed a greater emphasis on participatory approaches to dataset creation [48], which was used to some extent in building the ROOTS corpus [69], but more work is needed.

While full fledged community-driven datasets may seem impractical at the scale of pretraining data, approaches that elicit community input may aid in developing appraisal criteria for relevant problems like toxic language. Furthermore, novel technical approaches, such as modular and decentralized models [38, 53] may enable community-driven datasets, and deserve additional consideration. Yet 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 [68, 86].

An archival perspective demands that we recognize and confront the power dynamics at play in the creation 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.

5 Conclusion

The curation of pretraining data for LLMs has been largely attended to as an engineering exercise. Yet the curation of pretraining data is 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, towards considering their power (where pretraining data acts as an archive) and the processes that generated them. We highlight how common NLP practices 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. This framework offers a path forward to study not just the data, but the systems that produce it.

Acknowledgements We would like to thank Maria Antoniak, Kevyn Collins-Thompson, Irene Pasquetto, Jeremy Seeman, Amina Abdu, John Rudnik, and anonymous reviewers for their helpful feedback and suggestions.

References

[1] Aboriginal and Torres Strait Islander Data Archive (ATSIDA) (2013). ATSIDA protocols. `https://www.atsida.edu.au/protocols/atsida`.

[2] Bailey, J. (2013). Disrespect des fonds: Rethinking arrangement and description in born-digital archives. *Archive Journal*, 3:201–12.

[3] 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)*.

[4] 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*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[5] Brown, H., Lee, K., Mireshghallah, F., Shokri, R., and Tramèr, 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*, pages 2280–2292.

[6] 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. *Advances in Neural Information Processing Systems*, 33:1877–1901.

[7] Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., and Jurafsky, D. (2020). With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

[8] 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 *USENIX Security Symposium*, volume 6.

[9] Caswell, M. (2014). Toward a survivor-centered approach to records documenting human rights abuse: lessons from community archives. *Archival Science*, 14:307–322.

[10] Caswell, M. (2022). Inventing new archival imaginaries: Theoretical foundations for identity-based community archives.

[11] Chang, K. K., Cramer, M., Soni, S., and Bamman, D. (2023). Speak, memory: An archaeology of books known to ChatGPT/GPT-4. *arXiv preprint arXiv:2305.00118*.

[12] Chilcott, A. (2022). Towards protocols for describing racially offensive language in uk public archives. In *Archives in a Changing Climate-Part I & Part II*, pages 151–168. Springer.

[13] 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. *arXiv preprint arXiv:2204.02311*.

[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. (2011). *Documenting Society and Institutions: The Influence of Helen Willa Samuels*, pages 1–30. Society of American Archivists.

[16] Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. *Archival science*, 13:95–120.

[17] Cox, R. J. (1994). The documentation strategy and archival appraisal principles: a different perspective. *Archivaria*, 38.

[18] Davidson, S. B. and Freire, J. (2008). Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350.

[19] De Vynck, G. (2023). ChatGPT maker OpenAI faces a lawsuit over how it used people's data. *The Washington Post*.

[20] Debenedetti, E., Severi, G., Carlini, N., Choquette-Choo, C. A., Jagielski, M., Nasr, M., Wallace, E., and Tramèr, F. (2023). Privacy side channels in machine learning systems. *arXiv preprint arXiv:2309.05610*.

[21] 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).

[22] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[23] Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. (2019). Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

[24] 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*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[25] Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Pete Walsh, D. G., Soldaini, L., Singh, S., Hajishirzi, H., Smith, N. A., and Dodge, J. (2023). What's in my big data? *arXiv preprint arXiv:2310.20707*.

[26] 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 Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

[27] Fiesler, C. and Proferes, N. (2018). "Participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1).

[28] Flinn, A., Stevens, M., and Shepherd, E. (2009). Whose memories, whose archives? Independent community archives, autonomy and the mainstream. *Archival Science*, 9(1-2):71–86.

[29] Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

[30] 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*, pages 6786–6794, Marseille, France. European Language Resources Association.

[31] 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. *arXiv preprint arXiv:2101.00027*.

[32] Garcia1, G. G. and Weilbach, C. (2023). If the sources could talk: Evaluating large language models for research assistance in history. *arXiv preprint arXiv:2310.10808*.

[33] 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.

[34] 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*, pages 3356–3369, Online. Association for Computational Linguistics.

[35] Ghaddar, J. J. and Caswell, M. (2019). "To go beyond": Towards a decolonial archival praxis. *Archival Science*, 19:71–85.

[36] Gokaslan, A. and Cohen, V. (2019). Openweb-text corpus. `https://skylion007.github.io/OpenWebTextCorpus/`.

[37] 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*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[38] 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: Human Language Technologies*, pages 5557–5576, Seattle, United States. Association for Computational Linguistics.

[39] Haberstock, L. (2020). Participatory description: decolonizing descriptive methodologies in archives. *Archival Science*, 20(2):125–138.

[40] Ham, F. G. (1981). Archival strategies for the post-custodial era. *The American Archivist*, 44(3):207–216.

[41] Hedstrom, M. (2002). Archives, memory, and interfaces with the past. *Archival Science*, 2:21–43.

[42] 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.

[43] Hemphill, L., Schöpke-Gonzalez, A., and Panda, A. (2022). Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data*, 9(1):643.

[44] 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. *arXiv preprint arXiv:2207.00220*.

[45] 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*, 96(10.1029).

[46] Huvila, I. (2008). Participatory archive: Towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science*, 8:15–36.

[47] Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

[48] 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*, pages 306–316.

[49] Kandpal, N., Wallace, E., and Raffel, C. (2022). Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

[50] Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., Von Werra, L., Mou, C., González Ponferrada, E., Nguyen, H., et al. (2022). The BigScience ROOTS Corpus: A 1.6TB composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

[51] Lee, K., Cooper, A. F., and Grimmelmann, J. (2023). Talkin 'bout AI generation: Copyright and the generative-AI supply chain. *arXiv preprint arXiv:2309.08133*.

[52] 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 Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

[53] Li, M., Gururangan, S., Dettmers, T., Lewis, M., Althoff, T., Smith, N. A., and Zettlemoyer, L. (2022). Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*.

[54] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

[55] 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. *arXiv preprint arXiv:2305.13169*.

[56] McKemmish, S., Chandler, T., and Faulkhead, S. (2019). Imagine: A living archive of people and place "somewhere beyond custody". *Archival Science*, 19:281—301.

[57] Milligan, I. (2016). Lost in the infinite archive: The promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing*, 10(1):78–94.

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

[59] Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 79:119.

[60] O'Neal, J. R. (2015). "The right to know"': Decolonizing Native American archives. *Journal of Western Archives*, 6.

[61] OpenAI (2022). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

[62] Oren, Y., Meister, N., Chatterji, N., Ladhak, F., and Hashimoto, T. B. (2023). Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

[63] O'toole, J. M. (2002). Cortes's notary: The symbolic power of records. *Archival Science*, 2:45–61.

[64] 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(11):100336.

[65] 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*, pages 4296–4305, Online. Association for Computational Linguistics.

[66] 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 preprint arXiv:2306.01116*.

[67] Peng, K., Mathur, A., and Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922*, abs/2108.02922.

[68] 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*, pages 1–11.

[69] Piktus, A., Akiki, C., Villegas, P., Laurençon, H., Dupont, G., Luccioni, A. S., Jernite, Y., and Rogers, A. (2023). The ROOTS Search Tool: Data transparency for LLMs. *arXiv preprint arXiv:2302.14035*.

[70] 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. *arXiv preprint arXiv:2112.11446*.

[71] 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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[72] 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*.

[73] 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*, pages 840–854.

[74] Reardon, S., Samberg, R., and Vollmer, T. (2021). *Building Legal Literacies for Text Data Mining*, chapter 6, Ethics. University of California Berkeley Library.

[75] Richards, T. (1993). *The Imperial Archive: Knowledge and the Fantasy of Empire*. The Imperial Archive: Knowledge and the Fantasy of Empire. Verso Books.

[76] Roberston, A. (2023). ChatGPT returns to Italy after ban. *The Verge*.

[77] Samuels, H. (1986). Who controls the past. *The American Archivist*, 49(2):109–124.

[78] Samuelson, P. (2023). Generative AI meets copyright. *Science*, 381(6654):158–161.

[79] 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*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

[80] 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 *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

[81] Schellenberg, T. R. et al. (1956). *Modern archives*. University of Chicago Press Chicago, IL.

[82] 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(CSCW1).

[83] Schoenebeck, S. and Conway, P. (2020). Data and power: Archival appraisal theory as a framework for data preservation. *Proc. ACM Hum.-Comput. Interact.*, (CSCW2):1–18.

[84] Schwartz, J. M. and Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival Science*, 2:1–19.

[85] 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. *arXiv preprint arXiv:2310.16789*.

[86] Sloane, M., Moss, E., Awomolo, O., and Forlano, L. (2022). Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6.

[87] Small, Z. (2023). Sarah Silverman sues OpenAI and Meta over copyright infringement. *The New York Times*.

[88] Soldaini, L. (2023). AI2 Dolma: 3 trillion token open corpus for language model pretraining. *AI2 Blog*.

[89] 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(3):480–512.

[90] Stevens, M., Flinn, A., and Shepherd, E. (2013). New frameworks for community engagement in the archive sector: from handing over to handing on. In *Heritage and Community Engagement*, pages 67–84. Routledge.

[91] Subramonian, A., Yuan, X., Daumé III, H., 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: ACL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.

[92] The National Archives (2023). Cataloguing physical records: guidance for government departments.

[93] Thomas, D., Fowler, S., and Johnson, V. (2017). *The silence of the archive*. Facet Publishing.

[94] Trouillot, M. (1995). *Silencing the Past: Power and the Production of History*. Beacon Press books. Beacon Press.

[95] Underhill, K. J. (2006). Protocols for native american archival materials. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage*, 7(2):134–145.

[96] van Aken, B., Risch, J., Krestel, R., and Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

[97] Vitak, J., Shilton, K., and Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 941–953.

[98] 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*, pages 88–93, San Diego, California. Association for Computational Linguistics.

[99] 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. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

[100] 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 *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[101] Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, 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*, pages 4003–4012, Marseille, France. European Language Resources Association.

[102] Yakel, E. (2003). Archival representation. *Archival Science*, 3:1–25.

[103] Yakel, E. (2011). *Who Represents the Past? Archives, Records, and the Social Web*, pages 258–278. Society of American Archivists.

[104] Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., 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*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

[105] 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*, pages 19–27.

[106] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. F., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.