Disciplinary Practices in the Generation of Text Synthetic Data: A Critical Discourse Analysis

Adriana Alvarado Garcia[1], Nishanshi Atulkumar Shukla[2], Muneeza Azmat[1], Marisol Wong-Villacres[3]

[1]IBM Research, Thomas J. Watson Research Center
[2] UT Dallas
[3] Escuela Superior Politecnica del Litoral (ESPOL)
[1][adriana.ag, muneeza.azmat]@ibm.com, [2]nishanshi.shukla@utdallas.edu, [3]lvillacr@espol.edu.ec

Abstract

Synthetic data has emerged as an alternative or supplement to human-generated data, driven by several underlying assumptions that motivate its growing adoption among practitioners. These include the promise of increased efficiency by reducing the cost, time, and human labor involved in data collection and labeling, which is expected to potentially overcome data scarcity. Thus, as synthetic data becomes increasingly adopted to alleviate the data needs for Large Language Model development, it is critical to understand better the surrounding discourses and practices associated with their creation. We conducted a Critical Discourse Analysis on a corpus of 52 research articles from the Artificial Intelligence and Computational Linguistics conferences. As a result of our analysis, we identify three recurring disciplinary practices in establishing and reinforcing Cultural Scarcity and propose a set of recommendations to counteract it.

Introduction

The development of Artificial Intelligence (AI) and, more specifically, of Large language models (LLMs) have significantly transformed data practices. Since the performance of these technologies heavily depends on the scale, diversity, and quality of the data used to train and evaluate them (Liu et al. 2024), the demand for larger and more diverse datasets has significantly increased (Alzubaidi et al. 2023). However, human-generated data is often portrayed as inherently expensive and time-consuming to collect and annotate, a challenge exacerbated in specialized domains where data is scarce and human annotators with additional expertise are needed (Padmakumar et al. 2023; Muldoon et al. 2024; Kandpal and Raffel 2025). These challenges are further compounded by growing privacy regulations and ethical concerns, prompting the production of synthetic datasets as a practical solution to meet these escalating data requirements (Beduschi 2024).

Synthetic data refer to content or data generated artificially by algorithmic systems after mimicking real-world patterns (Jordan et al. 2022). Although synthetic data offers technical solutions to pressing challenges in AI development (Kumar 2024), it entails a host of inherent challenges regarding privacy, copyright infringement, and perpetuation of biases (Gal and Lynskey 2023; Lee 2024; Thakur and Hausenloy 2025; Nafis et al. 2025; Martino and Delmastro 2025), all suggesting a need to provide practitioners with particular forms of support.

In this work, we advance the understanding of such needed supports by unpacking the synthetic data generation practices that the AI discipline has institutionalized as acceptable and desirable, which we refer to as *disciplinary practices*. To identify and understand these disciplinary practices, we conducted a critical discourse analysis of 52 research articles from leading AI and computational linguistics conferences. Our work draws inspiration from Human-Computer Interaction (HCI) and Critical Data Studies research that argues that, to guide responsible dataset processes, it is critical to recognize these processes are not merely technical but shaped by practitioners' values, assumptions, and worldviews (Møller et al. 2020; Muller et al. 2019; Sambasivan et al. 2021; Orr and Crawford 2024; Qian et al. 2025; Alvarado Garcia et al. 2025).

Our analysis revealed three pervasive and interconnected disciplinary practices that, while technically sound, systematically impacted diverse states of the synthetic data pipeline: (1) operationalizing diversity, (2) repurposing existing data and technical artifacts, and (3) evaluating synthetic data with humans. These practices, we argue, contribute to what the French writer Jérôme Tharaud defined as *cultural scarcity* or the systematic removal or absence of cultural diversity, context, and alternative perspectives (Tharaud 2021), in this case, from synthetic data generation processes. We observed that cultural scarcity in this context manifested through the prevalence of hegemonic perspectives, the lack of diversified contexts, and the establishment of monolithic views of language, all contributing to synthetic data's failure to capture the rich complexity of human communication and experience. Further, our analysis indicates that cultural scarcity also shaped the practices we identified, thus creating self-reinforcing cycles that privilege dominant perspectives and marginalize alternative forms of knowledge. Such a perpetuation cycle has significant implications as synthetic data increasingly becomes the foundation for developing large language models and other AI systems that shape how technology mediates human interaction and understanding.

This paper makes three key contributions. First, we crit-

ically analyze AI disciplinary practices for generating synthetic data, revealing how seemingly neutral technical decisions embed cultural and political assumptions. Second, we use the concept of *'Cultural Scarcity'* as a framework for understanding how these practices systematically exclude diverse perspectives and contexts. Finally, we offer concrete recommendations for forms of support that help counteract cultural scarcity and develop more inclusive approaches to synthetic data generation.

Related Work

In recent years, the ease of generating synthetic data has led to its increasing use to the point that over 60% of the data used to develop AI and analytic projects is synthetically generated (Gartner 2023). Such a prevalence suggests that synthetic data will define the future of AI (Lee 2024). The generation process of synthetic data involves training models to capture patterns in real datasets, creating new data that does not map one-to-one to original sources (Tucker et al. 2020). Methodologies range from traditional statistical approaches to advanced deep learning techniques including diffusion models (Hao et al. 2024). Generative Adversarial Networks (GANs) have gained prominence through their adversarial training process involving generator and discriminator networks (Goodfellow et al. 2014). Large language models (LLMs) have also expanded possibilities for data augmentation and diversification in natural language processing tasks (Long et al. 2024).

Leveraging high-quality, conscientiously deployed synthetic data can offer a host of advantages to AI development. These include a significant reduction of the time-consuming, expensive, and laborious process of generating human-annotated data (Toews 2022; Lucini 2021), a limitless augmentation of scarce real-world data (Li et al. 2023; Wang et al. 2023), and the production of specifically-designed datasets that can ensure diversity and representativeness (Lee 2024). Such control and flexibility can enhance model performance and generalization across domains (e.g., health, privacy, red teaming) and formats (e.g., vision, audio, and text).

Despite these benefits, researchers in the fields of Machine Learning (ML), Human-Computer Interaction (HCI), and Critical Data Studies have emphasized that poorly deployed or low-quality synthetic data can pose significant issues. Jordon et al. (2022) noted that individuals' sensitive information can still be inferred from synthetic data, especially when combined with auxiliary datasets. Whitney and Norman (2024) highlighted "diversity-washing" and "consent circumvention" concerns even when synthetic data might appear to resolve distribution and representation criticisms. Synthetic datasets may also lack necessary ethical and legal constraints during creation, potentially replicating biases inherent to real-world data sources (e.g., internet data with societal biases) (Hao et al. 2024). Given the large scale of synthetic data, even small biases can lead to largely distorted outputs. The recursive training of LLMs, which entails AI systems generating synthetic data to train other AI systems for creating more synthetic data, can pose another potential harm by leading models to misperceive reality, also known as "model collapse" (Shumailov et al. 2024). Lastly, the gap in standardized metrics and evaluation frameworks for synthetic data (Long et al. 2024) can further complicate the aforementioned problems.

In response, diverse organizations have proposed responsible synthetic data generation frameworks. The United Nations University, for example, recommended guidelines addressing quality, security, and ethical implications (De Wilde et al. 2024), emphasizing transparency, accountability, and continuous improvement. Additional recommendations include establishing industry standards, documenting methods and parameters, and validating synthetic-trained models with real data. The Alan Turing institute published a report (Johansson et al. 2024) recommending differential privacy techniques, documenting data provenance, setting explicit use limitations, and establishing procedures for handling synthetic representations of harmful content.

Against this backdrop, various scholars have drawn from Science and Technology Studies (STS) to emphasize that guidelines and frameworks for mitigating AI harms must stem from an in-depth, critical understanding of the values encoded in the datasets sustaining AI technologies (Orr and Crawford 2024; Birhane et al. 2022; Peng, Mathur, and Narayanan 2021; Scheuerman, Hanna, and Denton 2021; Alvarado Garcia, Yang, and Miceli 2025; Miceli, Schuessler, and Yang 2020; Hutchinson et al. 2021). Specifically, they question AI datasets being neutral or universally beneficial and argue that datasets encode particular values, normative assumptions, and biases influenced by practitioners' personal decisions and institutional forces. As such, the call for dataset practitioners to be considered as "designers" of data: their perception, background, and motivation heavily shape the data they work with (Hutchinson et al. 2021; Chandhiramowuli et al. 2024; Feinberg 2017; Miceli, Schuessler, and Yang 2020). Understanding dataset practitioners' value-laden decisions, thus, can shed light on what the development of AI is prioritizing and working towards, the political and epistemological implications of these orientations, and the challenges and potential pathways forward (Birhane et al. 2022; Scheuerman, Hanna, and Denton 2021).

The emergent work coming from this body of research has focused on practitioners creating human-generated datasets only, highlighting that they often seek to meet the demands of clients and the market (Miceli, Schuessler, and Yang 2020) and, thus, favor the needs of research communities and large firms over broader social needs (Birhane et al. 2022). As such, their decisions tend to centralize power. Further, practitioners tend to justify their decisions through highly technical values (e.g., performance, efficiency, universality, impartiality, quantitative evidence, and novelty) (Scheuerman, Hanna, and Denton 2021; Birhane et al. 2022), struggle to articulate the ethical tradeoffs that their decisions could entail (Scheuerman, Hanna, and Denton 2021) (e.g., the reliance on shortcuts, lack of standardized practices, and use of ambivalent definitions of accountability for a dataset), and see the responsibility of such an articulation as outside their agency and capability (Widder and Nafus 2023). Considering these challenges, (Peng, Mathur, and Narayanan 2021) stresses that harm mitigation strategies

need to support dataset creators as well as users, researchers, and even conference program committees and provide support across datasets' entire life cycles.

Our work extends this line of research by critically exploring practitioners' decisions across the synthetic dataset generation process. Our goal is to inform mitigation strategies that acknowledge how the benefits and harms of synthetic data might be distributed unevenly and that challenge the priorities behind the increasing championing of synthetic data in the field of AI.

Methods

To examine the discourses and assumptions surrounding AI practitioners' disciplinary practices when generating synthetic data, we conducted a Critical Discourse Analysis (Jäger and Maier 2016) on a corpus of 52 research articles published at the five most prominent Artificial Intelligence conferences. We adopt Critical Discourse Analysis (CDA) as conceptualized by Siegfried Jäger (Wodak and Meyer 2015). Rooted in Foucault's theory of discourse, this approach to CDA distinguishes itself from other analytical methods by focusing on how power is embedded within texts. Thus, in this analysis, we understand discourse as a structured, institutionalized mode of speaking and writing that shapes and sustains action, thereby exerting power (Wodak and Meyer 2015). The *power of discourse* resides in its capacity to delineate the boundaries of what is sayable, producible, and perceptible within a given field of knowledge. In establishing these parameters, discourse simultaneously constrains or marginalizes alternative forms of knowledge that fall outside what can be articulated and materialized (Wodak and Meyer 2015). The following research question motivated our analysis: *What are the discourses and disciplinary practices guiding researchers in the process of generating synthetic data?*

Constructing the Corpus

To construct the corpus, we followed the guidelines of a systematic literature review (Moher et al. 2009). For this exploratory analysis, we decided to focus only on examining the practices that surround the generation of synthetic text data. Thus, to gather relevant research articles, we chose to restrict our search to two of the top conferences of the AI category on Google Scholar—NeurIPS and AAAI— and the three leading Computational Linguistics conferences—ACL, EMNLP, and NAACL—based on Google Scholar rankings. We searched for research articles throughout the existence of these conferences, i.e., NeurIPS from 1987-2023, AAAI 1980-2024, EMNLP from 1996-2023, ACL from 1979-2023, and NAACL from 2000-2024. In total, we looked at papers published in 165 iterations of conferences.

We use the term *"synthetic data"* within the abstract for the search query. In cases where abstracts of papers were not available directly on the website, we looked at the first page of .PDF file containing the paper [1]. As a result, we gathered a total of 756 papers. Since the focus of our research was on the generation of synthetic data, we further filtered these articles and selected those whose abstracts had keywords around the generation of synthetic data. These keywords include *"generat," "curat," "creat," and "taxonom"* These keywords were to account for different forms of words *generate, curate, create, and taxonomy* (e.g., generating, created, taxonomical, etc.). Through this process, we obtained a total of 406 papers. Then, the first and second authors independently annotated the abstracts of these 406 papers to 1) determine if they, in fact, described the process of generating synthetic data and 2) identify the type of data the articles focused on. Through this process, we arrived at a total of 82 papers. Lastly, the two first authors read the articles in their entirety and screened them using the following inclusion and exclusion criteria. We included articles with (1) an explicit section where the authors explained the process followed to generate text synthetic data, (2) a focus on developing tools, techniques or models for generating text synthetic data, (3) a focus on generating text synthetic data. We discarded articles that (1) described the use of synthetic data for evaluation of the central contribution of the article, (2) lacked a section of the process of generation of synthetic data or (3) focused on data other than textual data (e.g., tabular data). After this filtering process we gathered a corpus of 52 articles, being the oldest one from 2008 (See Fig.1 in the appendix for a detailed diagram of the process of building our corpus).

Data Analysis

The two first authors analyzed the 52 articles through various iterations over ten months. They started with the *structural analysis*, which consisted of reading the papers in full to identify relevant patterns and recurring themes. As a result of this initial stage, we identified an initial coding structure consisting of seven codes. This coding scheme helped us to clarify the papers' structure and facilitated the identification of practitioners' recurring assumptions and discursive strategies associated with generating synthetic data. We held weekly working sessions during the first coding phase, in which we discussed the codes and our emerging understanding of the corpus. Building on this initial structure, we revisited the articles of our corpus and identified typical discourse manifestations for each of the seven codes previously identified and refined the codes for the *detailed analysis*. Through this process, we began to record the authors' decisions in each paper, including the task of the focus of the paper, authors' decisions on reusing existing datasets or using new datasets, repurposing models for data generation, whether the authors conducted a human evaluation of the synthetic data, details on how the authors reported on annotators, authors' characterization of expertise, and discussion of disagreement. During this stage of the analysis, we also decided to quantify the trends we identified in the detailed analysis to determine how pervasive they were. Lastly, the synoptic analysis involved reviewing and integrating the findings from the structural and detailed analysis. During this phase, the two first authors engaged in multiple discus-

[1] For example, in the AAAI conference proceedings of the years 2008, 2007, 1998, 1997, 1993, 1991, 1990, 1988, and 1980, the abstracts were not accessible on the conference website, so we scraped the PDF of each paper to review the abstract.

sions to identify the predominant discourses and distilled the practices we present in the following section (See Table 1 in the appendix for a description of the codes from our analysis).

Findings: Disciplinary Practices

Generating synthetic data entails critical steps such as defining the problem that synthetic data will address and the sources to use as seed data, the generation of datasets, and the evaluation of the generated dataset. Through knowledge production, the AI community continuously defines the acceptable practices for conducting these steps. Understanding these practices—including how they take place and the assumptions enabling them—is vital for reflecting on addressing the potential risks of creating and using synthetic data.

Our critical discourse analysis of 52 papers describing different synthetic data generation processes emphasized three recurrent, interconnected practices critically impacting different steps of the synthetic data pipeline. First, addressing synthetic data problems as technological problems and, thus, working on them in disconnection with the particularities of the end user. As a result, the produced synthetic datasets are technologically feasible but have no clear applicability and impact in the users' context. Second, repurposing existing data artifacts to overcome data scarcity without considering their original purpose, bias, and other problems these artifacts might entail. This practice runs the risk of producing datasets that can perpetuate and augment critical issues that are hard to recognize or fix. Third, evaluating synthetic data with humans to ensure the adequacy of the end product but without consistent, standardized metrics, concepts, and approaches. This practice not only creates—and reinforces—a wide, confusing variation in what evaluation entails but can validate datasets that can entail critical, undetected problems for particular groups.

We found that these practices required papers' authors to navigate a significant tension between the care they sought to give to matters of richness, representation, and validity and the need to attain technically functional solutions. As a result, papers missed details and reflections about the impact of the decisions behind the produced synthetic data, creating a feedback loop that could reinforce problematic practices in synthetic data generation; they create an illusion of comprehensiveness that further legitimizes the missing details and reflection. We now describe each of the practices we identify while emphasizing the tensions they entail and the critical assumptions driving them. We analyze how each practice and its underlying assumptions create self-reinforcing cycles where reductive decisions, descriptions, and justifications prevail.

Practice 1: Operationalizing Diversity

The concept of diversity acknowledges the uniqueness and multiplicity of ways of knowing and existing in the world. Recognizing and aiming at attaining diversity in technology creation, thus, entails understanding and attending to the particularities of the context of use and the end-users (Zhao et al. 2024). Given the relevance of diversity, we analyzed how the papers of our corpus were operationalizing it to generate synthetic data. We found that about half of the articles recognized the concept's relevance (e.g., 25 papers from 52 papers in our corpus used the term *'diverse' or 'diversity'*). Further, eight of these papers directly addressed problems affecting non-dominant groups (e.g., the need for language-specific support) and worked on approaches to generate more diverse synthetic data, achieve better diversity, and analyze and evaluate diversity in training data. However, we found that most of these papers engaged rather superficially with diversity, operationalizing it as a variance of a general, context-agnostic problem. For example, they worked to provide a solution for speakers of a particular language without exploring those speakers' specific needs or experiences. Only one case from our corpus showed a user-centered operationalization of diversity that recognized particularities from the get-go. Such a prevalence of a context-agnostic, generalizable approach to diversity suggests that researchers tend to prioritize a generalizable computational solution over a solution that responds to contextual particularities. In what follows, we contrast both approaches to diversity and analyze the underlying assumptions that support them and their implications for the synthetic data generation process.

Diversity as the Response to a Clear End-User As we mentioned, we found only one case in the corpus we analyzed that approached diversity from an end-user perspective. This case described a novel approach to correct misspelled search queries in e-commerce settings. From the start, the authors specified an interest in supporting the specific context of users in regions with low English proficiency who often misspelled queries and, thus, struggled to produce successful results. As this quote exemplifies, the authors had a rich understanding of who they wanted to support, the nature and prevalence of their problem, and their limitations.

> *"An incorrectly spelt search query can return irrelevant products in e-commerce which hurts both the business and experience for a user who is unable to find the intended product. As per the latest English Proficiency Index report of written test data from 100 countries, 1 only ∼29% countries are proficient in English. Our platform operates in Asian countries like India which are ranked low in this survey. As per the latest Indian census only ∼10% of the Indian population is versed in the English language thus causing high spell errors in the user queries."*

To consider the end-user's characteristics—specifically, their English proficiency—the authors first categorized the errors these users were making. They explained their approach to attain this goal: "we describe various error classes based on the patterns we observe in our e-commerce search logs." As a result of their analysis, they were able to conclude that query reformulation alone was not an adequate solution for these users and their context. As they observed, such a solution would "fail to cover all kinds of errors like phonetic (cognitive) ones - "metras" vs "mattress" or edit/phonetic+word-compounding like "ball pen" vs "bolpan" when the user herself does not know the correct spelling. "

Establishing the error categories enabled the authors to recognize representation issues and, from there, define a particular goal for their synthetic data generation process: "We want frequent errors made by users to have a higher representation in our training data." This goal informed their next step: developing synthetic data sets for different spelling mistakes. To that end, they first carefully curated the seed data from logs.

> *"From our e-commerce search logs, we extract a seed set of clean (correctly spelt) queries on which we induce all types of errors. The clean queries are selected on the basis of their high volume and query CTR (Click Through Rate on the search result page). Errors are induced at word-level and then subsequently put back in the original query to generate the incorrect-correct training query pairs."*

From there, the authors created training pairs for a Transformer model and utilized weak supervision and curriculum learning to enhance the model's performance on complex spell corrections. As a result of the authors' careful operationalization of diversity as the integration of the end-users particular needs and realities along the synthetic data generation process, this study attained significant improvements in spell correction accuracy compared to existing models, ultimately improving both user experience and search result relevance on e-commerce platforms.

Diversity as a Variance of a General Problem The most prevalent approach to operationalize diversity that our analysis highlighted was treating it as a particular solution to a larger, more general problem. This, we observed, often entailed not engaging with the context where the produced synthetic data would be of use and instead approaching the problem from a general perspective. For example, authors of papers (Wu et al. 2016), who were addressing translation-related problems that involved non-dominant languages, approached the problem by choosing a standard form of a non-dominant language, thereby dismissing the different forms that a language could take (e.g., different dialects, accents, scripts) within already diverse, non-dominant realities.

For example, in their paper, (Wu et al. 2016) tackled the need for *"bilingually constrained synthetic data"* to *"alleviate the shortage of labeled data"*. To attain their goal, the authors extracted their BiSynData "from a combined corpus (FBIS and HongKong Law), with about 2.38 million Chinese-English sentence pairs" and generated 30,032 synthetic English and Chinese instances. The authors sought to include a diverse context by working with the Chinese language. However, they assumed that the Chinese language of all Chinese speakers was the same and that all these speakers had the same relationship with the English language. As such, they risked excluding cultural nuances from 'other' alternative forms of the same language.

In other cases, authors' dominant understanding of diversity came from a technical or theoretical perspective. For example, the authors in (Pratapa et al. 2018) presented a computational technique for creating artificial English-Spanish code-mixed (CM) language data. Once again, the authors addressed a diversity-related problem: code-mixed language is a form popular for bilingual speakers that entails using two or more languages in a sentence. However, by relying on the Equivalence Constraint Theory as critical guidance for generating CM sentences, the authors abided by a theoretically dominant perspective that they deemed "representative of real CM usage." As such, they treated the data in English and Spanish that they collected from sources such as Twitter as a variable input to a generalizable tool that could receive any other diverse input. In doing so, the authors ended up dismissing critical details of their decision, such as the distribution of audiences on Twitter, the possibility of the overarching presence of a dominant public, and the diverse cultures and backgrounds of the audiences generating these tweets (e.g., whose Spanish and English is being used and generated in code-mixed synthetic data? who does it? why? where? and which or whose form of this code-mixed language is being generated?)

Practice 2: Repurposing Existing Data and Technical Artifacts

Previous research has examined and cautioned against the uncritical reuse of datasets in machine learning, noting that when new models are built upon existing ones, the underlying datasets and model assumptions of the originals become embedded in the derived models, and assumptions that may no longer be appropriate and are often difficult to disentangle, especially when data are nested within opaque and black-box systems (Koch et al. 2021; Thylstrup et al. 2022; Park and Cordell 2023).

In our analysis, we identified that from our corpus, 47 research articles reported on reusing existing datasets, from which 32 articles contained data sourced from social media (e.g., Twitter), online platforms (e.g., Wikipedia), or media outlets (e.g., CNN). Only in five instances of our corpus, the authors reported on the creation of new datasets, of which, the data was extracted from user queries and interactions with authors' proprietary systems (n=2), directly extracted from the YouTube platform (n=1), from the 111th U.S. Congress bills (n=1), and from an agent-based (n=1). In addition to reusing existing datasets, we identify that in 46 articles, the authors reported on repurposing technical artifacts language models or large language models to generate synthetic data.

The authors relied on previously created datasets for a variety of reasons, the most recurring uses being *evaluation*, *model training*, and *generation of synthetic datasets*. Across our corpus, authors frequently used additional datasets for evaluation, primarily to assess the effectiveness of their frameworks and methods for generating synthetic data—such as list QA datasets, dialogues, or documents. These datasets also served as baselines to assess the utility and establish the quality of authors' generated synthetic data. Additional datasets also serve as benchmarks to evaluate the performance of the authors' proposed models to generate synthetic data.

When authors reused datasets for training purposes, they served various roles: as sources of context for generating synthetic questions and answers, as source domains for model training, and to train context generators for synthetic

data production. Lastly, when additional datasets were used for generating synthetic data, the authors of our corpus used extracts of the external datasets as *data seeds* [2] to generate synthetic training data.

We found that using data artifacts generally entailed overlooking essential aspects behind the repurposed artifacts, including the context of the artifacts' on-the-ground origins, the original purpose for creating this data, and the critiques behind the platforms that initially enabled the creation of these artifacts. Such overlooking suggests a high risk of embedding issues of representation and accountability in the produced synthetic data that could be significantly challenging to identify and address. Our analysis indicates that authors using data or technical artifacts without unpacking their origins and implications often focus only on overcoming data scarcity and, as such, overly rely on the research community's technological validation to choose and use an existing data solution. We illustrate these assumptions and their effects by describing how practitioners tend to repurpose these data and technical artifacts in more detail.

Overlooking the implications of modality transfer in data reuse Data modality refers to the mode in which data is collected, observed, or represented, such as text, audio, visual, or tactile data. A commonly overlooked implication of reusing datasets is transferring data collected in one modality for use in other modalities without adequately accounting for the consequences of missing the context of this transfer. In the analysis of our corpus, we identified that when practitioners resorted to repurposing datasets from unified multimodal experiences such as clinical interviews or therapy sessions and breaking them into discrete components to guide the generation of synthetic data, they overlooked how the deconstruction of datasets fundamentally altered what the data represents and means. Moreover, in following this approach, the authors tackled data purely from a technical perspective and applied a "divide and conquer" approach, removing each resulting piece from its original context. As a result, authors from our corpus inadvertently embedded their context-agnostic initial approach into their subsequent choices for generating synthetic data.

The case of (Chen et al. 2024) exemplifies this trend. In this paper, the authors describe an approach to detect depression through clinical interview transcripts. As the authors explained, automatic depression detection can be critical in contexts where professional clinicians are scarce. However, securing enough transcripts to train a model for such a task is not feasible due to privacy concerns. To augment real-world data, the authors decided to deconstruct clinical interviews transcripts, which are integrated, multimodal, and dynamic interactions, into discreet and analyzable components.

> *"Element Feature Extraction: In clinical interviews, each element serves as a unique indicator of the depressive state. Interview questions may probe for negative emotions, shedding light on underlying feelings of hopelessness or worthlessness. Answer transcripts may reveal a higher frequency of negative words, reflecting a depressed mindset. Answer audios can hint at depression through a flatter affect, slower speech rates, or prolonged pauses. Answer videos capture visual cues like reduced facial expressivity or subdued reactions. Together, these elements form a rich tapestry of information that helps in assessing an individual's depressive state."*

However, this modality transfer from integrated clinical interviews to independent decomposed entities dismisses critical implications. Clinical interviews derive their diagnostic value precisely from the interplay between these modalities: how a patient's verbal response relates to their tone, how their body language responds to specific questions, and how the emotions fluctuate in the interactions between doctor and patient. By treating each element of clinical interviews as an independent indicator, the authors inadvertently overlooked critical aspects behind clinical sessions, and disregard that depression assessment depends on understanding these elements as an integrated whole. These elements include essential contextual factors vital for the interpretation and sense-making process of assessing a person's mental state through an interview (e.g., the fluctuation of emotions in the interactions between doctor and patient, patients' prior mental health history, and the role of patients' cultural background in shaping emotions and responses). They also missed factoring in the professional training of those conducting the interviews. As a result, the produced dataset could, inadvertently, end up imposing a quite limited interpretation of the primary signals for assessing depression.

Our analysis highlighted that when authors repurposed data sourced from social media platforms, they missed another set of critical aspects related to these platforms. For example, to generate a *"Spanish therapy/counseling transcripts"* that could train chatbots mimicking mental health patients for learning purposes, the authors of (Lozoya et al. 2024) decided to compile *"30 hours of Spanish counseling session videos sourced from publicly available content on YouTube."* They did this as a response to the lack of data with *"valuable insights into various aspects [of therapy sessions], including cognitive patterns, interpersonal dynamics, and patient's goals and aspiration."* This modality transfer from embodied, visual therapy sessions to text-only synthetic data dismisses how therapeutic communication relies on non-verbal cues, spatial dynamics, and the embodied presence that cannot be captured in text alone. Overlooking the implications of modality transfer not only can strip away contextual meaning that arises from the interplay of different data modalities, but it also potentially leads to the generation of synthetic data that misrepresents how complex human experiences unfold. Like the previous case, the authors failed to consider critical details embedded in these data (e.g., the mental health issues discussed) and not recorded in it (e.g., patients' socio-economic and cultural background and the therapists' level of expertise). Further, they missed the impact that the platform's aspects could have on the resulting

[2]Data seeds refer to the initial examples or pieces of data used as starting points from which language models can generate new instances while maintaining specific characteristics or patterns.

synthetic dataset, such as the authenticity of the YouTube data and the justification behind the selection criteria for the videos used.

Dismissing the implications of repurposing publicly-available data The work of (Shi, Chen, and Zhao 2024) exemplifies the implications of repurposing publicly-available data. In this work, the authors reported *"a novel pipeline for automatically constructing large-scale preference data"* to *"construct the SAFER-INSTRUCT (SI) dataset, a safety preference dataset for LLMs"*. To construct this dataset, the authors used *"datasets widely available in the NLP community"* on the categories of hate speech, self-harm content, sexual content, and illegal activities. These datasets, however, were sourced from social media data for a different purpose than safety prevention and did not always abide by the authors' definitions of the categories they were working on. For example, the authors defined the self-harm category as *"content that encourages performing acts of self-harm, such as suicide, cutting, and eating disorders, or that gives instructions or advice on how to commit such acts.*. However, they used the SCDNL dataset to inform it without considering that this dataset's goal is actually to *"address the unexplored issue of classifying between depression and more severe suicidal tendencies using web-scraped data and neural networks"*.

By repurposing this dataset to generate safety-focused instructions, the authors ran the risk of generating safety-focused instructions that unintentionally conflated acts of self-harm with depressive and suicidal behaviors. This mix-up of definitions has larger implications for how LLMs and synthetic data represent knowledge: as these conflations continue happening over time, the knowledge that LLMs produce could become diluted, and tracing back definitions and decisions could become impossible. Further, this case demonstrates that an underlying assumption of those creating synthetic data is that all publicly available data that touch on topics of interest is generalizable to address the need for more data production.

The produced synthetic data could also embed problematic platform-specific elements. For example, when the authors generated synthetic data for safety, they missed reflecting on critical questions such as "How is this data defining safety?" and "Safety for whom?" In generating a dataset containing preferred and dispreferred responses for unsafe instructions, the authors overlooked that social media platforms like Twitter (now X) and specialized subreddits represent specific audiences from certain privileges. The absence of critical reflection on such data and instructions generated based on that leads to a hegemonic perspective of safety as this task focuses only on safety for a particular audience.

Overlooking the gaps and biases of LLMs Within the articles of our corpus 46 papers used a language model or an LLM to generate synthetic data. In total, the authors from our corpus reported using 32 models for data generation, the most used of which were the BART and GPT -2 models. Given the extensive pre-trained knowledge, strong language understanding, and instruction-following abilities of LLMs, they have become a practical substitute and complement to data created by humans (Budnikov, Bykova, and Yamshchikov 2025; Long et al. 2024).

While adopting language models or large language models for data generation has become an increasingly accepted practice within the AI community, the fact that the generated data is derivative from the base knowledge of the model used for its generation imposes inherent challenges, such as data gaps and biases of the generator models (Whitney and Norman 2024; Gallegos et al. 2024), which are challenging to address due to the lack of transparency of the datasets used to trained LLMs (Hardinges, Simperl, and Shadbolt 2024; Schaul, Yu Chen, and Tiku 2024). Thus, it is critical for researchers using these data artifacts to reflect on the potential effect that the use of specific LLMs can have on the synthetic data produced.

As the following quote exemplifies, some of the authors in our corpus did recognize the limitations of LLMs in regards to biases, transparency, and stability of results. *"As our method relies on large pretrained language models, it should be noted that users deploying these technologies need to be aware of their undesirable, human-like biases (Sheng et al., 2019; Abid et al., 2021). Methods for reducing these harmful associations are actively being developed by the research community (Liang et al., 2021; Schick et al., 2021)."*

However, we did not find deeper reflections on the potential gaps that using LLMs to generate synthetic data could produce nor critiques around the opacity of these data artifacts. Papers often described LLMs as a quite promising and unique data artifact that have demonstrated *"exceptional performance across a broad range of NLP tasks [3, 39, 28, 37, 38, 58]"* and *"the ability to generate highly fluent and coherent textual data"* which use was specifically ideal *"for tasks where high-quality datasets are not readily available or access to real data is limited or expensive"*. Even when other papers described LLMs in a less positive light, they emphasized how the impact is still unknown, thus, prioritizing the feasibility of use over the problems such a use might entail.

"Reinforcing LM biases. A point of concern for the authors is the unintended consequences of this iterative algorithm, such as the amplification of problematic social biases (stereotypes or slurs about gender, race, etc.). Relatedly, one observed challenge in this process is the algorithm's difficulty in producing balanced labels, which reflected models' prior biases. We hope future work will lead to better understanding of the pros and cons of the approach."

Our analysis of paper (Chen et al. 2024)—which we initially presented at the beginning of this section —illuminates the problems that these different ways of positioning LLMs within the synthetic data generation process might cause. While the authors of this article turned to LLMs for augmenting clinical interviews data, the resulting synthetic dataset could face critical challenges to represent the variation of the interviews' answers needed. Ideally, diversity in responses for this case should highlight the vast spectrum of human emotions and mental states, such as those experiencing depression, anxiety, high levels of stress, or a more stable emotional state. They should also provide responses that reflect patients unique ways of thinking and experiencing the

world, and in their responses, there would be diversity in the tone, speed, and level of deepness and details, which are beyond a mere variation in vocabulary. Without a detailed description of how the LLMs produces synthetic datasets, it becomes impossible to shed light on why the end result tackles or not these critical aspects.

Practice 3: Evaluating Synthetic Data with Humans

In our analysis, we identified that out of the 52 papers of our corpus, in over half of them (n=32), the authors decided to include humans in evaluating the generated synthetic data. While we did not find papers justifying this decision, existing research positions human evaluation as the golden standard (more about that here please). We observed, however, significant inconsistencies in how each article addressed critical aspects of these human-based evaluations, potentially undermining the reproducibility and comparability of results. Specifically, we found a lack of standardization in the selection, definition, and level of justification of (i) evaluation goals, (ii) metrics, and (iii) human participation. The pervasiveness of such inconsistencies suggests that researchers prioritize the presence of humans in the evaluation process over ensuring that the process takes place with standardized definitions.

Pursuing Customized Evaluation Goals In the articles of our corpus, we observed high levels of variation in how authors defined the same aspects of their evaluation goals. In particular, they used different definitions of adequacy for synthetic data and approaches to determine such adequacy. They also made different sampling-related decisions. Across articles, however, we found no justification for these differences.

We observed a lack of a standardized definition for synthetic datasets' adequacy and, thereby, of approaches to determine adequacy. In regards to pursuing adequacy, oftentimes authors referred to it as the degree to which the synthetic data resembled human-generated content, often described as "human likeness" or "resemblance to real data" (i.e., data generated by humans). However, authors defined 'human likeness' differently, and their definitions were often ambiguous. For instance, one article described it as a composite of *"the fluency, coherence, and engagement of the response, i.e., whether it resembles a human response"*. Another defined it as *"the perceived resemblance of the synthetic transcripts to real therapy transcripts"*.

Regarding evaluation approaches, they ranged from rating the synthetic data on a scale to binary classification to attribute-specific scoring according to predefined criteria. Once again, these approaches often relied on the authors' definitions of quality, which varied in each study. For example, one article asked evaluators to rate the output based on a four-level rating system that the authors had implemented "for categorizing the quality of models' outputs." Another article producing synthetic therapy transcripts asked a group of eight psychologists to rank the transcripts' resemblance to real therapy transcripts.

Another discrepancy we observed was in the decisions about sampling the data for evaluation. While some authors would ensure human participants had ample time to evaluate the data and, thus, assigned them *"no more than 10 examples per rater,"* others would assign them 1000 items *"to classify if a question is in-domain or out-of-domain."*

Using Multiple but Undefined Metrics Authors in our corpus frequently relied on ad hoc evaluation approaches, applying a range of metrics without a shared or standardized definition. Our analysis identified six recurring dimensions used to assess synthetic data: quality, consistency, coherence, fluency, relevance, and correctness. Notably, no study evaluated the data using a single metric in isolation. Instead, authors consistently combined multiple dimensions, reflecting a fragmented yet thoughtful effort to capture the complexity of synthetic data evaluation. For example, *"humans evaluated the synthetic overlap summaries along the four dimensions: Coherence, Consistency, Fluency, Relevance"*.

We also observed that the definitions of the metrics most frequently used were not standardized. For example, to evaluate summaries produced, (Amplayo, Angelidis, and Lapata 2021) defined coherence as the answer to the question, "is the summary easy to read and does it follow a natural ordering of facts?" In the case of (Bansal, Akter, and Karmaker Santu 2022), they resorted to existing literature to operate under a quite detailed definition of coherence:

> *"Coherence: It represents the collective quality of all sentences. This dimension aligns with the DUC quality question of structure and coherence whereby the generated summary/document should be well-structured and well-organized. It should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic"*

Offering undetailed descriptions of human participation As studies in the fields of ML and Critical Data Studies have stressed, documenting the perspectives of those annotating or evaluating data is critical to understanding possible biases (Miceli, Schuessler, and Yang 2020; Guest et al. 2021). However, in our corpus, the description of the human participants responsible for evaluating synthetic data also varied widely. Table 2 in the appendix, for example, shows how 17 out of the 32 papers that conducted human evaluation did not report basic details about the annotators' demographics, training, background, or identity. In cases where such information was included, it was often minimal. For example, one study noted that the data were *"rated on a 1–5 scale for each criterion by one expert and reviewed by another"*. At the same time, another described *"asking three domain experts to rate a question as good or bad based on four attributes"*.

As these examples highlight, the characterization of human participants' expertise also varied greatly: although both studies referenced the use of experts, they did not define the criteria for this designation. Further, across the 52 papers analyzed, only 7 explicitly used the terms *experts* or *domain experts* to refer to individuals involved in generating or evaluating synthetic data. In those cases, definitions and justifications of expertise also varied significantly.

Only 3 of the 7 papers provided specific details about the experts' training, identifying them as clinical psychologists, trained speech-language pathologists (SLPs), or *"NLP researchers, each with at least one year of specific experience in the task of factual consistency evaluation"*. The remaining four papers exhibited inconsistencies in how expertise was defined and justified, limiting their explanation regarding expertise to statements such as *"these datasets comprise evidence texts from biomedical literature with manual annotations by experts"* and *"For supervised data, we use PQA-L(abeled) subset of 1K question-passage pairs manually curated by domain experts."*

Cultural Scarcity and How to Counteract It

Our analysis of 52 papers from leading venues in AI research identified three practices that the AI discipline follows to generate synthetic data: (1) operationalizing diversity, (2) repurposing existing data and technical artifacts, and (3) evaluating synthetic data with human input. Further, in line with previous studies on real-world dataset creators, we found that synthetic data practitioners conducted these practices primarily focusing on technical values such as universality, performance, and efficiency, often missing to reflect on the social implications of their work.(Birhane et al. 2022; Scheuerman, Hanna, and Denton 2021; Widder and Nafus 2023). Our study advances these findings by describing how these values affected synthetic data generation. Specifically, we illustrate how the ease and speed of synthetic data generation amplify the *cascade effect* (Sambasivan et al. 2021), thereby propagating unexamined social issues in data and AI systems more quickly and massively. Moreover, we emphasize that given practitioners' repeated reuse of problematic data and technical artifacts, tracing and addressing the root of these issues is almost impossible,

Drawing on this reflection, we argue that the AI discipline currently follows practices to generate text synthetic data that can lead to what Tharaud defined as *cultural scarcity*. That is, the systematic elimination of local *"knowledge, languages, and cultural resources"* such that traditional forms of subjectivity are lost (Tharaud 2021, 413-414). In synthetic data, perpetuating cultural scarcity entails systematically removing cultural diversity, context, and alternative perspectives in the seed data and the synthetic data generation process, thus losing subjectivity and context and widely accepting a *"view from nowhere."* But, as scholars of critical data studies have argued, data always possess a *"view from somewhere"* (Jones 2022), and even when overlooked, context is always present (Seaver 2015). In this sense, cultural scarcity in synthetic data generation refers to normalizing dominant views and contexts by overlooking whose perspectives and worlds are being represented, generated, and ignored. Building on Tharaud's analysis of the power of cultural scarcity (Tharaud 2021, 413-414), we argue that cultural scarcity in synthetic data generation manifests in three forms: prevalence of hegemonic perspective, lack of diversified contexts, and establishment of monolithic languages. In the following section, we first describe how the three practices we identified contribute to distinct manifestations of cultural scarcity. While we present each practice as individual decisions, they never occurred in isolation; instead, they were intertwined and reinforced each other. Then, we offer a set of recommendations and potential directions for future research aimed at counteracting cultural scarcity.

Multiple Manifestations of Cultural Scarcity

Reinforcing hegemonic perspectives: Hegemonic perspectives refer to the dominance of a particular worldview and way of life in shaping public perceptions of reality. It involves controlling cultural and ideological knowledge production in favor of dominant groups (Altheide 1984). Our analysis stressed that, in the case of synthetic data generation, the dominance of a hegemonic perspective began when practitioners engaged in the discretionary work of translating high-level objectives into tractable, supposedly context-neutral problems. Previous data science research has examined the complexities of making these difficult translations, highlighting the need for subjective decisions throughout the process and emphasizing that each practical choice carries social and ethical implications (Passi and Barocas 2019; Hutchinson et al. 2021). While we acknowledge the inherent challenges practitioners faced in reinterpreting tasks to design suitable synthetic datasets, our analysis revealed that, in doing so, they often overlooked diversity and disregarded the specific contextual needs where the synthetic data would ultimately be applied.

Our analysis also highlighted that repurposing existing tools and datasets can further reinforce hegemonic perspectives. As we saw, relying on existing datasets perpetuated hegemonic viewpoints by obscuring earlier subjective judgments and prioritizing over-represented dominant cultural narratives. (Park and Jeoung 2022) stressed the need to understand further the potential social impacts of reusing machine learning datasets within existing metadata schemas. Our analysis suggests that social implications could be further exacerbated by the practice of dismissing information about the human annotators who assessed the quality and utility of data. It is complex to discern which perspectives are privileged in the datasets without knowing who these annotators are.

Stripping away context and disregarding diversity: Our study illustrated that cultural scarcity emerged when practitioners engaged in practices that removed context and dismissed the diversity of synthetic data's potential users. As shown in *Practice 2,* relying on existing datasets significantly contributes to the lack of diversified contexts. Datasets collected from available digital platforms systematically underrepresent cultural contexts with less digital presence or resources for data collection and curation. Over the years, researchers have pointed to the limited representation of communities across platforms depending on geographies and internet access (Tufekci 2014; Olteanu et al. 2019), and the potential risks of making conclusions from such sources (boyd and Crawford 2012; Alvarado Garcia, Yang, and Miceli 2025). When practitioners reused datasets, they neglected to consider the distribution of people and perspectives and assumed that the large scale of datasets would ensure results that were diverse enough to represent all possi-

ble voices and experiences. Similarly, by repurposing LLMs for synthetic data generation, practitioners further perpetuated cultural scarcity as these models, trained predominantly on accessible Western corpora, have limited exposure to diverse knowledge systems and epistemological frameworks (Whitney and Norman 2024). Lastly, given that we ignore which expertise and perspectives are included during the human evaluation of synthetic data, it is even more challenging to determine which contexts and perspectives are included and which ones are dismissed to determine the utility and quality of synthetic data.

Promoting a monolithic view of language: We refer to a monolithic view of language as the notion that languages exist in a singular and uniform form, excluding variations such as dialects. In essence, the monolithic view of language does not represent the reality of languages. In our corpus, we observed how practitioners' decisions led them to contribute to a monolithic view of language consistently. First, when practitioners determined the kind of data they needed by formulating the tasks they would work on as context-agnostic, they inadvertently stripped away linguistic particularities and cultural idioms in favor of generalized formulations that privileged dominant language patterns. Then, when practitioners repurposed existing data and technological artifacts, they further contributed to the entrenchment of a monolithic language: existing datasets often overrepresented dominant languages and dialects while marginalizing others and LLMs trained on such data reproduced the hegemonies embedded in their corpora, thereby homogenizing linguistic diversity even when producing superficially varied outputs. Lastly, the systematic lack of details on human evaluations created additional filtering mechanisms where evaluators, typically unfamiliar with marginalized languages, applied quality judgments that favored dominant linguistic structures.

Recommendations to Counteract Cultural Scarcity

Support the design of synthetic datasets: Practitioners require structured support to identify the types of data necessary for their tasks and to avoid challenges similar to those we described in *Practice 2* related to the clinical interview settings. Such support may involve, for instance, a critical examination of what is lost in the translation of task formulations, which is an essential step in determining appropriate data requirements. Domain experts could provide this support. Our analysis showed a widespread lack of expert involvement across the various stages of synthetic data generation. This often left practitioners to make ad hoc decisions without adequate guidance. Once practitioners undermine the domain-specific knowledge essential to carry out a task, they unintentionally embed their assumptions, values, and subjectivities about what constitutes specialized practices into making synthetic datasets, which may compromise its validity. The importance of involving domain experts in data annotation processes to improve accuracy (Guest et al. 2021) and to develop an empirical understanding of the knowledge gap between experts and non-experts(Waseem and Hovy 2016) has been acknowledged by previous work. We argue that the role of expertise should start from the earlier stages of the synthetic data generation lifecycle and call for the development of tools and methodologies that enable domain expert knowledge to help practitioners determine the data types necessary for specific tasks.

Support the documentation of synthetic data generation: Even though data documentation within the AI and ML research communities has received considerable attention (Gebru et al. 2021; Paullada et al. 2021; Miceli et al. 2021, 2022), none of the articles in our corpus provided evidence that the synthetic data generation process was documented in any substantive way. While the decisions made during synthetic data generation have significant implications, the lack of clear documentation makes it hard to understand their consequences. In the past, traditional human-generated datasets were often accessible for auditing and determining, up to a certain extent, their impact on the performance of AI technologies. However, given the current practices of generation and documentation of synthetic data, such transparency has mainly become absent because there is often no clear record of how the data were designed, what sources informed their generation, or the assumptions underlying their construction. And even if we knew, as we showed in our analysis, the fact that practitioners rely on existing datasets and additional models for the synthetic data creation further complicates identifying what exactly goes into the generation of synthetic datasets. Thus, we recommend developing tools and mechanisms that facilitate the documentation of the decisions involved in generating synthetic data. Such documentation tools could include the definition of seed data and the selection of models for data generation. They could even include statistics, both from whom the data were collected and of those who contributed to labeling or annotating the data, to prevent overgeneralization and the exclusion of certain demographic groups during reuse, as previous research has suggested (Hovy and Spruit 2016).

Support the identification of data gaps: Practitioners need to actively interrogate which communities, perspectives, and knowledge are being dismissed when reusing datasets and repurposing technical artifacts to generate synthetic data. In our analysis, we traced back the original purpose and category definitions of the datasets that articles in our corpus reused as seed data. Through this analysis, we observed that when practitioners reused datasets, they not only dismissed the original purpose behind the dataset creation but, in some instances, reassigned meaning to them. Preserving context is just as crucial as maintaining content in data curation, as it enables data reusers to decide whether to adopt existing data or gather their own (Faniel, Frank, and Yakel 2019). However, current community incentives do not encourage this practice, and there is also a lack of available tools to record how data are being repurposed as seed data. Therefore, drawing from previous recommendations (Park and Jeoung 2022; Park and Cordell 2023), we consider that there is a need to develop tools and mechanisms that support practitioners in evaluating datasets' quality, relevance, and original context before reusing them for synthetic data generation, helping ensure that such reuse aligns with their intended purpose and mitigates potential harm.

References

Altheide, D. L. 1984. Media Hegemony: A Failure of Perspective. *Public Opinion Quarterly*, 48(2): 476–490.

Alvarado Garcia, A.; Candello, H.; Badillo-Urquiola, K.; and Wong-Villacres, M. 2025. Emerging Data Practices: Data Work in the Era of Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Alvarado Garcia, A.; Yang, T.; and Miceli, M. 2025. What Knowledge Do We Produce from Social Media Data and How? *Proc. ACM Hum.-Comput. Interact.*, 9(1).

Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A. S.; Al-dabbagh, B. S.; Fadhel, M. A.; Manoufali, M.; Zhang, J.; Al-Timemy, A. H.; and et al. 2023. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1).

Amplayo, R. K.; Angelidis, S.; and Lapata, M. 2021. Unsupervised Opinion Summarization with Content Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14): 12489–12497.

Bansal, N.; Akter, M.; and Karmaker Santu, S. K. 2022. Learning to Generate Overlap Summaries through Noisy Synthetic Data. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11765–11777. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Beduschi, A. 2024. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data & Society*, 11(1): 20539517241231277.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–184.

boyd, d.; and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5): 662–679.

Budnikov, M.; Bykova, A.; and Yamshchikov, I. P. 2025. Generalization potential of large language models. *Neural Computing and Applications*, 37: 1973–1997.

Chandhiramowuli, S.; Taylor, A. S.; Heitlinger, S.; and Wang, D. 2024. Making data work count. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–26.

Chen, Z.; Deng, J.; Zhou, J.; Wu, J.; Qian, T.; and Huang, M. 2024. Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8181–8194. Mexico City, Mexico: Association for Computational Linguistics.

De Wilde, P.; Arora, P.; Buarque de Lima Neto, F.; Chin, Y.; Thinyane, M.; Stinckwich, S.; Fournier-Tombs, E.; and Marwala, T. 2024. Recommendations on the Use of Synthetic Data to Train AI Models.

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

Feinberg, M. 2017. A design perspective on data. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2952–2963.

Gal, M.; and Lynskey, O. 2023. Synthetic Data: Legal Implications of the Data-Generation Revolution. *Iowa Law Review*, 109: Forthcoming. LSE Legal Studies Working Paper No. 6/2023.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.

Gartner. 2023. Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning — gartner.com. https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning. [Accessed 23-05-2025].

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

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1336–1350. Online: Association for Computational Linguistics.

Hao, S.; Han, W.; Jiang, T.; Li, Y.; Wu, H.; Zhong, C.; Zhou, Z.; and Tang, H. 2024. Synthetic data in AI: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*.

Hardinges, J.; Simperl, E.; and Shadbolt, N. 2024. We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/xau9dza3.

Hovy, D.; and Spruit, S. L. 2016. The Social Impact of Natural Language Processing. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598. Berlin, Germany: Association for Computational Linguistics.

Hutchinson, B.; Smart, A.; Hanna, A.; Denton, E.; Greer, C.; Kjartansson, O.; Barnes, P.; and Mitchell, M. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 560–575. New York,

NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Jäger, S.; and Maier, F. 2016. Analysing discourses and dispositives: A Foucauldian approach to theory and methodology. *Methods of critical discourse studies*, 109–136.

Johansson, P.; Bright, J.; Krishna, S.; Fischer, C.; and Leslie, D. 2024. Exploring responsible applications of Synthetic Data to advance Online Safety Research and Development. *arXiv preprint arXiv:2402.04910*.

Jones, C. 2022. The view from somewhere: Critically reflexive recruitment heuristics for big data. *First Monday*, 27(2).

Jordan, J.; Houssiau, F.; Cherubin, G.; Cohen, S. N.; Szpruch, L.; Bottarelli, M.; Maple, C.; and Weller, A. 2022.

Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S. N.; and Weller, A. 2022. Synthetic Data–what, why and how? *arXiv preprint arXiv:2205.03257*.

Kandpal, N.; and Raffel, C. 2025. Position: The Most Expensive Part of an LLM should be its Training Data. arXiv:2504.12427.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research.

Kumar, A. 2024. The Role of Synthetic Data in Advancing AI Models: Opportunities, Challenges, and Ethical Considerations. *Journal of Artificial Intelligence General Science*, 5(1): 443–459.

Lee, P. 2024. Synthetic Data and the Future of AI. *Cornell Law Review*, 110(Forthcoming).

Li, Z.; Zhu, H.; Lu, Z.; and Yin, M. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10443–10461. Singapore: Association for Computational Linguistics.

Liu, Y.; Cao, J.; Liu, C.; Ding, K.; and Jin, L. 2024. Datasets for Large Language Models: A Comprehensive Survey. arXiv:2402.18041.

Long, L.; Wang, R.; Xiao, R.; Zhao, J.; Ding, X.; Chen, G.; and Wang, H. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11065–11082. Bangkok, Thailand: Association for Computational Linguistics.

Lozoya, D.; Berazaluce, A.; Perches, J.; Lúa, E.; Conway, M.; and D'Alfonso, S. 2024. Generating Mental Health Transcripts with SAPE (Spanish Adaptive Prompt Engineering). In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5096–5113. Mexico City, Mexico: Association for Computational Linguistics.

Lucini, F. 2021. The Real Deal About Synthetic Data — sloanreview.mit.edu. https://sloanreview.mit.edu/article/the-real-deal-about-synthetic-data/. [Accessed 23-05-2025].

Martino, F. D.; and Delmastro, F. 2025. Challenges and Limitations in the Synthetic Generation of mHealth Sensor Data. arXiv:2505.14206.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–25.

Miceli, M.; Yang, T.; Alvarado Garcia, A.; Posada, J.; Wang, S. M.; Pohl, M.; and Hanna, A. 2022. Documenting Data Production Processes: A Participatory Approach for Data Work. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Miceli, M.; Yang, T.; Naudts, L.; Schuessler, M.; Serbanescu, D.; and Hanna, A. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 161–172. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Moher, D.; Liberati, A.; Tetzlaff, J.; and Altman, D. G. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, 151(4): 264.

Møller, N. H.; Bossen, C.; Pine, K. H.; Nielsen, T. R.; and Neff, G. 2020. Who does the work of data? *Interactions*, 27(3): 52–55.

Muldoon, J.; Cant, C.; Wu, B.; and Graham, M. 2024. A typology of artificial intelligence data work. *Big Data & Society*, 11(1): 20539517241232632.

Muller, M.; Lange, I.; Wang, D.; Piorkowski, D.; Tsay, J.; Liao, Q. V.; Dugan, C.; and Erickson, T. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15.

Nafis, N.; Esnaola, I.; Martinez-Perez, A.; Villa-Uriol, M.-C.; and Osmani, V. 2025. Critical Challenges and Guidelines in Evaluating Synthetic Tabular Data: A Systematic Review. arXiv:2504.18544.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2: 13.

Orr, W.; and Crawford, K. 2024. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, 26(9): 4955–4972.

Padmakumar, A.; Inan, M.; Gella, S.; Lange, P.; and Hakkani-Tur, D. 2023. Multimodal Embodied Plan Prediction Augmented with Synthetic Embodied Dialogue. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6114–6131. Singapore: Association for Computational Linguistics.

Park, J.; and Cordell, R. 2023. The ripple effect of dataset reuse: Contextualising the data lifecycle for machine learning data sets and social impact. *Journal of Information Science*, 0(0): 01655515231212977.

Park, J.; and Jeoung, S. 2022. Raison d'être of the benchmark dataset: A Survey of Current Practices of Benchmark Dataset Sharing Platforms. In Shavrina, T.; Mikhailov, V.; Malykh, V.; Artemova, E.; Serikov, O.; and Protasov, V., eds., *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, 1–10. Dublin, Ireland: Association for Computational Linguistics.

Passi, S.; and Barocas, S. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 39–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

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.

Peng, K.; Mathur, A.; and Narayanan, A. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922*.

Pratapa, A.; Bhat, G.; Choudhury, M.; Sitaram, S.; Dandapat, S.; and Bali, K. 2018. Language Modeling for Code-Mixing: The Role of Linguistic Theory based Synthetic Data. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1543–1553. Melbourne, Australia: Association for Computational Linguistics.

Qian, C.; Liu, M. X.; Reif, E.; Simon, G.; Hussein, N.; Clement, N.; Wexler, J.; Cai, C. J.; Terry, M.; and Kahng, M. 2025. LLM Adoption in Data Curation Workflows: Industry Practices and Insights. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Schaul, K.; Yu Chen, S.; and Tiku, N. 2024. Inside the secret list of websites that make ai like chatgpt sound smart.

Scheuerman, M. K.; Hanna, A.; and Denton, E. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–37.

Seaver, N. 2015. The nice thing about context is that everyone has it. *Media, Culture & Society*, 37(7): 1101–1109.

Shi, T.; Chen, K.; and Zhao, J. 2024. Safer-Instruct: Aligning Language Models with Automated Preference Data. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7636–7651. Mexico City, Mexico: Association for Computational Linguistics.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759.

Thakur, M.; and Hausenloy, J. 2025. Opportunities and Challenges of Frontier Data Governance With Synthetic Data. arXiv:2503.17414.

Tharaud, J. 2021. Western Salvage: Scarcity, Settler Colonialism, and Adaptation in Wallace Stegner's Wolf Willow. *ISLE: Interdisciplinary Studies in Literature and Environment*, 30(2): 406–425.

Thylstrup, N. B.; Hansen, K. B.; Flyverbom, M.; and Amoore, L. 2022. Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, 9(2): 20539517221139785.

Toews, R. 2022. Synthetic Data Is About To Transform Artificial Intelligence — forbes.com. https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/. [Accessed 23-05-2025].

Tucker, A.; Wang, Z.; Rotalinti, Y.; and Myles, P. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1): 147.

Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 505–514.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Andreas, J.; Choi, E.; and Lazaridou, A., eds., *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: Association for Computational Linguistics.

Whitney, C. D.; and Norman, J. 2024. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1733–1744.

Widder, D. G.; and Nafus, D. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620.

Wodak, R.; and Meyer, M., eds. 2015. *Methods of critical discourse studies*. Introducing Qualitative Methods. Sage, 3rd edition. ISBN 9781446282410.

Wu, C.; Shi, X.; Chen, Y.; Huang, Y.; and Su, J. 2016. Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2306–2312. Austin, Texas: Association for Computational Linguistics.

Zhao, D.; Andrews, J. T. A.; Papakyriakopoulos, O.; and Xiang, A. 2024. Position: measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.