Position: Cracking the Code of Cascading Disparity Towards Marginalized Communities

Golnoosh Farnadi [1] **Mohammad Havaei** [1] **Negar Rostamzadeh** [1]

Abstract

The rise of foundation models holds immense promise for advancing AI, but this progress may amplify existing risks and inequalities, leaving marginalized communities behind. In this position paper, we discuss that disparities towards marginalized communities – performance, representation, privacy, robustness, interpretability and safety – are not isolated concerns but rather interconnected elements of a *cascading disparity phenomenon*. We contrast foundation models with traditional models and highlight the potential for exacerbated disparity against marginalized communities. Moreover, we emphasize the unique threat of cascading impacts in foundation models, where interconnected disparities can trigger long-lasting negative consequences, specifically to the people on the margin. We define marginalized communities within the machine learning context and explore the multifaceted nature of disparities. We analyze the sources of these disparities, tracing them from data creation, training and deployment procedures to highlight the complex technical and socio-technical landscape. To mitigate the pressing crisis, we conclude with a set of calls to action to mitigate disparity at its source.

1. Introduction

Foundation models with their ability to learn and adapt across various domains, are rapidly transforming the landscape of AI. However, these large-scale models that often trained on massive, unfiltered datasets, pose various risks for marginalized communities. Foundation models can perpetuate and amplify existing biases, leading to disparities in performance, privacy, robustness, model understanding, and even the generation of harmful content for marginalized communities. For example, large language models (LLMs) often perform worse on language tasks involving dialects spoken by low-resource languages (Liang et al., 2021). An image generation or diffusion model primarily associating certain professions with specific genders or ethnicities (Luccioni et al., 2023). Multimodal models can struggle with recognizing or classifying images of people from marginalized groups, particularly those with darker skin tones or features that do not align with dominant beauty standards (Buolamwini & Gebru, 2018; Schwemmer et al., 2020). A voice assistant models often misinterpret commands spoken with a regional accent which leads to frustration and exclusion for the user (Tatman, 2017). Multimodal models might generate hateful or offensive content that perpetuate discrimination and incite violence against already vulnerable groups (Zellers et al., 2019).

While extensive research has focused on identifying and addressing specific types of disparities (see Section 3), a holistic understanding of how these disparities are interconnected remains largely unaddressed. This narrow focus can worsen existing biases and introduce new ones, disproportionately harming marginalized communities. For example, counterfactual data augmentation (Gardner et al., 2020) that have shown promising out-of-domain generalizability (Samory et al., 2021), if implemented without careful consideration, can reinforce harmful stereotypes or lead models to violate established social norms (Sen et al., 2022).

Moreover, standard evaluation benchmarks, such as StereoSet (Nadeem et al., 2020), often focus on surface-level assessments of foundation models using non-robust prompting metrics or post-deployment downstream task evaluations. These approaches fail to directly probe the deeper sources of bias embedded within the models.

In this position paper, we argue that these disparities are interconnected elements of a **cascading disparity phenomenon** affecting marginalized communities. We discuss that representational disparities within the model lie at the root of this phenomenon. The distinct, complex distributions representing marginalized communities are often insufficiently captured during training, leading to the ”flattening” of their representations within the model. This, in turn, manifests as performance disparities, reinforcement of stereotypes, privacy violations, and other types of harmful disparities. By analyzing the sources of disparities throughout the lifecycle of foundation models, from data collection

[1]Google Research, Montreal, Canada. Correspondence to: Golnoosh Farnadi <gfarnadi@google.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

and training to adaptation and deployment, we highlight the complex technical and socio-technical landscape that shapes this problem. We urge researchers to move beyond conventional loss function optimization when training foundation models. It's vital to develop metrics that directly assess the quality of representations with regards to marginalized communities, ensuring that the model learns the nuanced, low-dimensional manifolds associated with these groups. Additionally, we advocate for investigating how a model's capacity should be dynamically allocated across different distributions to achieve a more equitable representations. Our contributions in this position paper are as follows: 1) Identifying and categorizing disparities in foundation models that disproportionately impact marginalized communities (Section 3). 2) Defining the cascading disparity phenomenon and how it stems from representational disparities (Section 4). 3) Analyzing the origins of disparities across the lifecycle of foundation models (Section 5). 4)Providing a list of call to actions to address representational disparities in foundation models (Section 6).

2. Marginalized Communities

Before discussing how and why the current way of training and deploying models creates disparity towards marginalized communities, it's important to clearly define what we mean by marginalized communities or "data at the margin" in the context of machine learning (ML). Marginalized communities (Allman, 2013) or as Williams and White put it "marginalized from mainstream society" (Williams & White, 2003) refers to groups systematically excluded and discriminated against based on factors like race, ethnicity, gender, sexual orientation, socioeconomic status, disability, religion, or other identity aspects (Allman, 2013; Williams & White, 2003). This historical exclusion often results in ongoing lack of representation, resources, and ongoing discrimination. Such historical exclusion is reflected and amplified in ML applications. Data from marginalized communities is often missing, underrepresented, or misrepresented in training datasets. This leads to several data-centric challenges for ML models: 1) **Small sample size:** Marginalized communities are underrepresented in datasets. 2) **Disparate distribution:** The data distributions associated with marginalized communities may differ significantly from the majority population. This can encompass factors like demographics, language use, or behavioral patterns. 3) **Complex distributions:** Data may exhibit nuances and complexities due to intra-group diversity, cultural patterns, or unique historical contexts.

These factors create challenges for ML models, making it difficult to represent marginalized communities accurately. In statistical terms, these low-sample classes form the "long tail" of distributions. While this typically refers to low-occurrence events, in this context, it highlights data samples rarely seen during training, despite their real-world prevalence not necessarily being lower than more common data classes.

Note that this definition focuses primarily on the data-centric aspects of marginalization in machine learning and it's essential to acknowledge that definitions of marginalized communities are not solely technical but also inherently social and political. In the context of technology and socio-technical systems, marginalized communities can be understood as groups of people who experience: i) *Systemic Disadvantage*: These communities face historical, social, political, and economic barriers that limit their access to opportunities, resources, and power. This systemic disadvantage often stems from factors like discrimination, prejudice, and social exclusion. ii) *Data Exclusion and Invisibility*: Marginalized communities may be underrepresented or even invisible within data sets used to train and develop technological systems. And iii) *Limited Agency and Participation*: Marginalized communities may have limited opportunities to participate in the design, development, and deployment of technological systems that significantly impact their lives.

3. Types of Disparity

In this section, we present how foundation models systematically disadvantage marginalized communities through the following eight disparities.

Embedding/Representation disparities Foundation models serve as powerful tools for representation learning with the goal of automatically capturing meaningful and generalized features from the data that makes it easier to extract useful information in down stream tasks. A good representation is one that captures the underlying explanatory factors for the observed input. As such representation learning is closely linked to manifold learning with the hypothesis that high dimensional data lies on a low dimensional manifold (Gorban & Tyukin, 2018). It is generally understood that learning complex features leads to better generalization in downstream tasks (Bengio et al., 2013; Natekar & Sharma, 2020). A representation should be expressive enough to capture complexities expressed through factors of variation for every subgroup in of the input space. With that notion, representation disparity is defined as constrained complexity in learned embeddings due to data limitations, resulting in challenges in manifold creation. Prior work has either focused on removing sensitive attributes with adverserial debiasing (Zhang et al., 2018) and contrastive learning (Tian et al., 2020) or maintaining semantic distances in the embedding space (Zafar et al., 2017; Beutel et al., 2017; Zhang et al., 2018), addressing the limited complexity issue.

Performance disparities Performance disparity is defined as disparities in model performance between majority and minority populations in downstream tasks. Previous work has shown such performance gap is manifested in healthcare (Hall et al., 2023), text summarization (Yang et al.,

2023), translation (Prates et al., 2020), image classification (Ali et al., 2023), and recommender systems (Moradi & Farnadi, 2023). Extensive research has explored the performance gap, with various fairness metrics like demographic parity, accuracy parity, equal opportunity, and equalized odds (Hardt et al., 2016) reflecting these disparities for marginalized communities.

Privacy disparities Privacy disparity is defined with a propensity for memorization more pronounced for marginalized communities (Carlini et al., 2022a). Limited model capacity results in prioritized generalization for larger populations, exacerbating privacy concerns (Tramèr et al., 2016; Carlini & Wagner, 2018) and catastrophic forgetting (Luo et al., 2023). Existing work have shown privacy-enhancing technologies such as differential privacy in SGD (DP-SGD) (Abadi et al., 2016) that rely on gradient clipping and noise injection, disproportionately degrade accuracy of marginalized communities (Bagdasaryan et al., 2019; Malekmohammadi et al., 2024). Furthermore, model compression techniques on foundation models such as iterative magnitude pruning (Maene et al., 2021), which can result in enhancing overall privacy of the model , proportionately impact on the accuracy of communities at the margin (Hooker et al., 2020; 2019; Tran et al., 2022; Hashemizadeh et al., 2023)

Robustness disparities Robustness disparity is the variation of the performance, accuracy and reliability of the ML model across different populations that could particularly impact the marginalized communities. This issue arises from inadequate representation of these communities in various stages of ML development, including (i) data creation (ii) model development and (iii) deployment. On *data* side, marginalized communities are usually out-of-distribution samples. During the learning, their distribution is often miss-represented and under specified. This would make marginalized data also more prone to adversarial and poisoning attacks due to the lack of generalization of the model for these groups (Madry et al., 2017; Athalye et al., 2018; Ma et al., 2022). Finally during the deployment process, conditions that the model is deployed for the marginalized communities are often dismissed and they lack proper testing for edge cases, group-specific perturbations and robustness testing towards factors of variation within these groups, that leads to increased vulnerability to adversarial attacks and failures in marginalized populations.

Hallucination[1] **disparities**. It is widely acknowledged that Large Language Models (LLMs) and by extension Visual Language Models (VLMs) suffer from Hallucinations. These instances manifest as the model confidently generates output that while seeming plausible, are unreasonable or factually untrue with respect to the source of information. While the source of hallucination is not yet fully understood, hallucination in LLMs typically arise from the inherent data limitations in the training data and complexity of the model architecture (Ji et al., 2023; Dziri et al., 2022). Hallucination disparity is defined as an elevated likelihood of generating fabricated or hallucinated outputs for marginalized communities due to data limitations. Wang & Sennrich (2020) showed that hallucinations are more prevalent for out-of-domain distributions compared to in-domain distributions for Neural Machine Translation. Cohen et al. (2018) also demonstrated that a mismatch in distribution between source and target domains in image translation causes the model to hallucinate confounding factors when generating samples from the target domain.

Insufficient information and memorization tendencies contribute to the model making erroneous assumptions about the data and consequently leading to higher likelihood of generating inaccurate outputs for marginalized communities (Wang & Sennrich, 2020; Cohen et al., 2018; Arjovsky et al., 2019; Guo et al., 2018). These erroneous assumptions and misrepresentations manifest themselves in the learned manifold of the foundational model. From this perspective, hallucination disparity is closely linked to representation disparity, wherein the model has not acquired expressive representations of marginalized groups. The model's failure to capture the subtleties inherent to these groups within its learned representations is a contributing factor to the emergence of hallucinations in its generated outputs.

Note that while hallucination gap can be categorized under the broader category of performance disparity, we believe highlighting hallucinations as a distinct issue allows us to emphasize the importance of addressing outputs that are factually incorrect or misleading. This has significant implications for model reliability and factfulness, that warrants focused attention beyond the performance disparity in downstream tasks that are often focused on supervised learning tasks with existing measures such as demographic parity or equalized odds.

Model Understanding disparities Foundation models, with their vast number of parameters and complex training data, are often challenging to fully comprehend. The lack of explainability is amplified when the model's training data is not representative of diverse populations (Du et al., 2020). Trying to understand a model that generates text with underrepresented groups might lead to inaccurate assumptions or explanations that are not truly reflective of how the model works. This can result in misinterpreting the outputs and assigning incorrect attributions to the model's behavior. Since model decisions are based on countless data points and parameters, determining the specific reasons for a particular output is often difficult (Zhao et al., 2023). This becomes even more challenging when a model has not been exposed to sufficient data or diverse perspectives regarding marginalized groups. Without deep knowledge of a model's inner workings and data, there is a risk of simplifying its behavior. Moreover, the people developing the models and interpreting will be less likely to identify issues related to

[1]Also referred to as "Confabulation" in literature

marginalized groups. Such oversimplifications may overlook the nuances or complexities involved when the model interacts with topics related to marginalized communities.

Model Multiplicity/Underspecification disparities Due to model uncertainty, the predictions or text generated by the model can be seemingly arbitrary or random when addressing topics related to marginalized groups (Black et al., 2022). The model might generate responses that are off-topic, insensitive, or even harmful. Foundation models often involve stochastic (random) processes during the generation of text. This randomness, combined with a lack of understanding of underrepresented groups, can amplify the arbitrary nature of the outputs, making them increasingly unpredictable (Ganesh et al., 2023; D'Amour et al., 2022). The stochastic elements in these models can sometimes exaggerate the biases present in the training data. This can lead to the generation of random outputs that inadvertently amplify stereotypes or misinformation concerning marginalized communities, Ganesh showed that due to model multiplicity, the random behavior of the model is higher for marginalized groups (Ganesh, 2024).

We intend to underscore how model design limitations (e.g., architectures that are too broad or too narrow) can specifically lead to ambiguity and uncertainty in model behavior. This connects model architecture choices directly to issues of bias and fairness towards marginalized communities. Our classification system aims to draw attention to these specific nuances within the broader *performance disparity* category. This will facilitate more targeted analysis of existing mitigation strategies, even while acknowledging the interconnected nature of these issues.

Safety disparities Regarding the detection and mitigation of safety concerns, conflicts could arise due to differing moral values and cultural contexts between groups (Scherrer et al., 2023; Benkler et al., 2023). Existing work show that current models may reflect dominant Western cultural biases and values (Rao et al., 2023). E.g., A model used to detect hate speech may have difficulty identifying slurs or harmful language directed towards certain groups due to underrepresentation in its data (Deshpande et al., 2023). Similarly, an AI model used to write text or stories may generate content that reinforces stereotypes or reflects biases against marginalized groups if it has limited exposure to inclusive data (Blodgett et al., 2020).

4. Cascading Disparity: A Systemic Issue of Interlinked Disparities due to Embedding Disparity

The eight disparities that we discussed in the previous section, while distinct in nature, are not independent issues. They interact and reinforce each other, creating a cumulative negative impact on marginalized communities. In this section, we show that at the root of this complex issue lies embedding disparity that its influence cascades exacerbating other forms of disparity.

Imbalances in how foundation models represent various groups within their embedding space directly contribute to performance disparities in downstream tasks. When models lack expressive representations of marginalized groups, their performance suffers in tasks that involve those groups. Such lack of proper manifold creation for marginalized communities and their underrepresentation, also force the model to use it's capacity to memorize specific data points instead of learning generalizable representations. Marginalized communities are often underrepresented which can make their data points seen as outliers, and make them more susceptible to privacy attacks under the privacy onion effect, outlined by Carlini et al. (Carlini et al., 2022b). If the model memorizes specific data points (common for marginalized groups), its behavior is inconsistent and challenging to even explain or interpret.

Models with poor embedding representations of marginalized groups are also less robust, leading to higher uncertainty. Limited representation of marginalized groups in the embedding space leads to unpredictable and arbitrary predictions due to increased sensitivity to minor input changes. And while hallucinations can occur for various reasons, they often stem from model uncertainty. When a model is less certain about how to handle data from a marginalized group, it is more likely to fabricate or "hallucinate" details that are not grounded in the data.

The lack of manifold embedding disparity in foundation models is a critical safety concern. If marginalized groups are misrepresented in the embedding space, the model may fail to recognize their unique perspectives, cultural contexts, needs, and languages specific to those communities (Jha et al., 2024; Qadri et al., 2023). This lack of representation can result in outputs that overlook the concerns of marginalized groups or provide inaccurate or inappropriate responses to their needs, effectively excluding them from the benefits these models could offer or even perpetuate harmful stereotypes, reinforce exclusion, and amplify hate speech. When a model fails to learn the patterns associated with diverse perspectives, it struggles to generalize its knowledge to unseen scenarios or when presented with prompts related to underrepresented groups. This limitation results in unpredictable and uncertain outputs. Insufficient and imbalanced embedding space can cause the model to associate certain attributes more strongly with some groups than others. This might result in inconsistent outputs, where the same prompts produce different responses depending on the perceived identity of the subject. This uncertainty can lead to outputs that are biased or discriminatory.

We intentionally focused on eight well-established categories of responsible AI to demonstrate the interconnectedness of disparities. We highlight how overlooked intersections can magnify harms, alongside widening perfor-

mance gaps. To mitigate the cascading negative impacts on marginalized communities, we need to address these interlinked disparities at the core. Next, we discuss the sources of manifold embedding disparities in foundation models.

5. Sources of Disparity

While both traditional ML models and foundation models can exhibit disparities, the nature and sources of these disparities can differ significantly. Foundation models, due to their scale and complexity, present unique challenges in terms of disparity identification, mitigation, and societal impact. While the previous section highlighted the concerning disparities that can arise with foundation models, understanding the sources of these disparities is crucial for effectively addressing them. In this section, we explore how and why current practices in foundation model development can amplify and perpetuate harmful disparities, particularly against marginalized communities, and how they can differ from traditional models in several key ways.

5.1. Design and Data Collection

One fundamental challenge in training ML models, including traditional and foundation models, revolves around the obstacles posed by data. Here, we argue how and why data issues are contributing to representation disparity. Traditional ML models, typically require smaller, and domain-specific datasets tailored to the specific task, and the disparities often arise from the specific data used to train the model. If the data is imbalanced or contains inherent historical disparities, the model will likely learn and perpetuate those disparities. In foundation models, due to their massive scale and reliance on diverse datasets, they can be susceptible to a wider range of data biases. Data used to train foundation models can be orders of magnitude larger than those used in traditional ML, which presents challenges in data storage, processing, and ensuring data quality. Uneven or non-random sampling methods can lead to datasets that underrepresent or misrepresent certain demographics, creating skewed data distributions that disadvantage marginalized communities (Passi & Barocas, 2019). The accessibility and availability of data can also vary across different groups (Olteanu et al., 2019). Moreover, societal biases ingrained in cultural norms, and historical data can be inadvertently embedded within datasets, leading to models that reflect and amplify these biases (Pedreschi et al., 2009; Richardson et al., 2019). Finally, differences in how individuals interact with technology, i.e., digital gap (Hargittai, 2011), or provide data (Olteanu et al., 2019) can introduce biases into datasets. For instance, marginalized communities may have limited access to technology or may be hesitant to provide data due to privacy concerns, leading to models that are less accurate or perform poorly on tasks involving these groups (Molamohammadi et al., 2023)

5.2. Training Procedure and Learning Algorithm

Disparities in traditional models are typically studied through the data or the chosen loss function. However, in foundation models, due to their complex learning processes and interactions with vast amounts of data, they can exhibit disparities that were not present in the training data or algorithms. These emergent disparities can be difficult to anticipate and address. Bellow, we discuss the source of these disparities during the training process that although they can occur in traditional models, the magnitude of the issue in foundation model can differ significantly. Sara Hooker (Hooker, 2021) discussed how we should look beyond data to discuss disparities in ML models and consider the choices that we make, e.g., the algorithms or hyperparameters, to study algorithmic discrimination. Bellow, we extend her analysis and discuss the impact of various choices that we make during training that can have a significant impact on the outcome:

Loss function: Foundation Models diverge significantly from traditional ML models in terms of their loss function design. Foundation models often employ self-supervised learning paradigms that emphasize learning the sequential structure of data instead of directly optimizing for performance on a specific task (Brown et al., 2020; Devlin et al., 2018). A common example of this is the ”next token prediction” objective, where the model is trained to predict the next word or token in a given sequence. This shift towards sequence learning has resulted in remarkable improvements in the language capabilities of large language models, among other areas. In traditional supervised ML, loss functions are designed to identify features and patterns in the data that are strongly correlated with a particular label or target variable. The model learns to prioritize the features most relevant to the task at hand. However, foundation models trained with sequence prediction objectives learn representations that capture the sequential dependencies and underlying structure of the data, irrespective of specific labels. This enhances their ability to generalize to a diverse range of downstream tasks without the need for extensive task-specific fine-tuning (Radford et al., 2019). However, this focus on sequential learning has a profound impact on how foundation models process and learn from data, e.g., the pre-training data for the model might reflect biases related to the order or sequence in which information is presented such as sequential exposure to gendered language can reinforce gender stereotypes (Bolukbasi et al., 2016; Sun et al., 2019). Existing work indicates that in scenarios where there exists a non-linear relationship between group membership (e.g., considering demographics like race or gender) and a specific outcome, using a single linear classifier often leads to a performance trade-off. One, or perhaps both, of the groups involved will experience a decline in model performance (Dwork et al., 2018). This is because linear classifiers, by nature, struggle to capture the complexity of these non-linear relationships. In certain

cases, however, incorporating information about the group differences directly into the design of the machine learning model can lead to the development of simpler learned functions (mathematical representations) that ultimately enhance the performance across various groups (Dwork et al., 2018; Suresh & Guttag, 2019). By understanding the nuances of the group differences, models can be tailored to better learn the diverse patterns within the data. However, such mitigation cannot simply be transferred to foundation models due to the complexity of identifying marginalized groups.

Aggregation Method: The aggregation method for calculating the overall loss across samples influences how the model learns the representation of different groups within data. When data points are aggregated with similar weights, the model tends to prioritize learning the distribution of larger populations. If a loss function overemphasizes majority groups, existing work show that this leads to under-representation or misrepresentation of marginalized groups (Suresh & Guttag, 2019; Mehrabi et al., 2021). This results in the model neglecting the complexities of marginalized communities, often leading to performance gaps and a tendency towards memorization rather than generalization in representing these groups (as discussed in Section 3).

Data order: The order in which data points are presented during training, particularly in the context of marginalized groups, can significantly impact the model's performance (Ganesh et al., 2023). When data points are read randomly, there is a risk of catastrophic forgetting (Kirkpatrick et al., 2017), that the model will tend to forget the patterns associated with marginalized groups encountered earlier in the training process. This oversight often results in suboptimal model performance for these groups. However, research into the effects of fine-tuning has demonstrated that strategically positioning marginalized group data toward the end of the training process can improve their representation in the model (Dodge et al., 2020). This is because, in the fine-tuning stage, the model has a chance to reinforce its understanding of the marginalized group's patterns while the knowledge of the majority groups remains relatively stable, potentially leading to better performance for the marginalized communities.

Batch size: In the presence of random data selection, the aggregation of gradients from various data samples also plays a crucial role in determining the model's gradient norms and update directions. Due to the inherent distributional differences between marginalized communities and majority groups, gradient updates from these groups are likely to conflict, potentially in terms of both their directions and magnitudes (Suresh & Guttag, 2019). When the training dataset comprises a significantly larger proportion of data from the majority groups, the model's overall learning trajectory is likely to be dominated by these majority data points. As a result, the learning of patterns from marginalized groups, effectively their 'voices', can be suppressed during the training process.

Batch Composition: Similar to the considerations of batch size and data presentation order, the composition of the training batch can also exert significant influence on the model's behavior, particularly when it comes to learning about marginalized communities. If smaller batch sizes are employed, and these batches disproportionately consist of data from a marginalized community, there is a higher likelihood that the aggregated gradient updates from these batches will notably influence the model's learning trajectory, making it more receptive to the patterns and characteristics present in the marginalized group's data.

Learning Rate: Learning rate and training duration exhibit a disproportionate influence on error rates on marginalized communities specifically those on the long tail of the distribution. Studies on deep neural network memorization demonstrate a delayed learning process for the marginalized communities (Jiang et al., 2020). Consequently, a common early stopping approach can carry the potential to systematically bias performance against certain data distributions (Hooker, 2021).

5.3. Deployment and Adaptation

One of the challenges of traditional supervised models are their adaptability to a new task or domain. Traditional ML models are often designed and trained from scratch for specific tasks. This focused training while beneficial for performance on targeted task, restricts their transferability to new tasks and domains. Transfer learning approaches try to address this limitation by leveraging knowledge from source tasks to target tasks and domains without the need to train from scratch. These approaches could be Homogeneous (Zhuang et al., 2020; Weiss et al., 2016), when the feature and label spaces remain consistent across domains but differ slightly in their distributions. Homogeneous techniques aim to minimize these distribution discrepancies. On other other hand, transfer learning approaches could be Heterogeneous techniques (Zhuang et al., 2020; Weiss et al., 2016) when labels, and feature spaces differs and they aim to bridge the gaps between different distributions and feature spaces.

Creation of the foundation models created a significant boost to the improvement of transferablity of the knowledge between domains, and tasks. However, unlocking the potential of foundation models for real-world applications requires effective adaptation to specific downstream tasks. There are two primary adaptation paradigms for foundation models: **fine-tuning** and **prompt engineering**. In this section, we explore their distinct strengths, and limitations, and discuss how each approach can potentially mitigate or amplify the disparities towards marginalized community.

5.3.1. Fine-tuning

One of the most widely used transfer learning approaches, that is usually a homogeneous approach, is fine-tuning (Brown et al., 2020). Fine-tuning offers a means to

tailor foundation models to specific tasks and domains by adjusting internal parameters based on task-specific data and domain knowledge. This can significantly improve performance and achieve high accuracy for downstream tasks which requires specialized knowledge. However, fine-tuning poses several challenges such as catastrophic forgetting (Kirkpatrick et al., 2017). Minimizing the loss for adjusting the pre-trained model to a new task or data, could substantially drop the performance on some of the original seen data.

RLHF (Reinforcement Learning from Human Feedback) is a broader adaptability technique compared to traditional fine-tuning that is only optimizing the model for limited tasks and domains. RLHF is a multifaceted training approach that refines the model behavior with human preferences and feedback (Bai et al., 2022). RLAIF (Reinforcement Learning from AI Feedback) is similar to RLHF, however, it leverages another AI model to automatically generate feedback on the outputs of the base model being trained (Lee et al., 2023). Note that although these models could be aligned with some human values and feedback, it can also inherit the biases of the people whose feedback is involved in the model after fine-tuning.

5.3.2. Prompt Engineering

The training process for fine-tuning can be computationally expensive. Hence, instead of extensive fine-tuning, recent and popular paradigm of adaptation of foundation models is though well-crafted prompts that can guide foundation models to generate desired outputs or perform specific tasks without additional training. Popular prompt engineering and in context-learning approaches suggest that instead of extensive fine-tuning, well-crafted prompts can guide foundation models to generate desired outputs or perform specific tasks without additional training.

Although recent in-context learning efforts attempt to debias foundation models (Dwivedi et al., 2023), these approaches have limitations. Since in-context learning does not modify model parameters, which would fundamentally increase the model's representational capacity, post-processing techniques alone cannot fully address representation disparity or alter the model's core understanding of marginalized communities. While in-context learning can influence model behavior through text conditioning, and safeguards can mitigate specific biases or manage harmful responses, these strategies offer limited bias reduction and cannot fundamentally change the model's inherent representations issues.

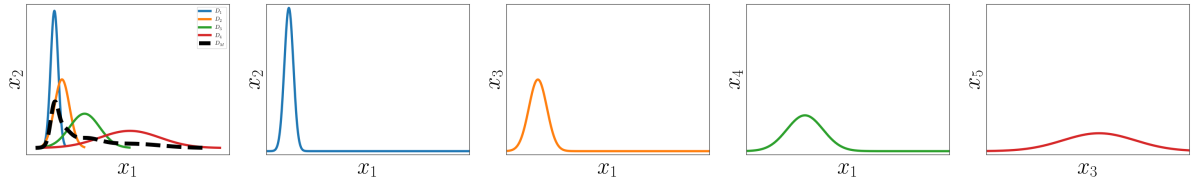
6. Technical Gaps & Call To Actions

As outlined earlier, marginalized communities experience multifaceted disparities rooted in the lifecycle of foundation models. Here, we mainly focus on the technical limitations that hinder efforts to address these issues. We will demonstrate how training methodologies, despite relying on high-quality data, can reinforce cascading disparity phenomena and we propose a set of call for actions to mitigate the root of such disparities, i.e., representation disparity. We acknowledge that the calls for action in this position paper are mainly technical, and we note other sociotechnical dimensions of disparities that are important to notice but out of the scope of our position paper in the impact statements.

Call to Action: Developing Guidelines for Training under Mixtures of Heterogeneous Distributions: First, we must emphasize that the data used to train foundation models often originates from diverse underlying distributions. One of the fallacies in training foundation models is the simplifying assumption that the underlying distribution is a long-tailed distribution (see Figure 1, sub-figure (a), D_m). While long-tailed distributions (like those in recommender systems due to popularity bias) are a relevant factor, specifically in traditional ML models, the disparities facing marginalized communities stem not just from limited data points, but from fundamental differences within their distributions.

As we defined in Section 2, we account for three characteristics for marginalized community, size, distinct distribution and distribution complexity. Hence, to mitigate representation disparity, this distinction with long-tailed distribution and the consideration of mixture of distribution is essential to grasp (see Figure 1, sub-figure (a)). Furthermore, even considering the mixture of distributions assumption, methods are needed to address high-dimensional datasets that inherently exist across multiple low-dimensional manifolds (see the left plot on Figure 1). We argue that considering low-dimensional manifolds to learn representation of each distinct distribution can significantly improve the representation of marginalized communities and reduce the representation gap. One could consider hierarchical manifold learning similar to hierarchical Bayesian models to capture global and local variances or dependencies, such as Bayesian meta learning (Ravi & Beatson, 2018).

While increasing model capacity could theoretically help learn under mixtures of heterogeneous distributions, it is important to consider practical limitations (i.e., learning all the underlying dimensions x_1, x_2, x_3, x_4, x_5 in Figure 1). The computational complexity of blindly pouring data into the training would quickly drive up costs, making it an unsustainable solution in real world applications. Here, we speculate less computational and resource extensive solutions to address the representation disparity by fixing the data ordering and batch aggregation problems discussed in section 5. We suggest to strategically organizing data points based on their underlying distributions. Grouping data from similar distributions minimizes conflicting gradient signals in training. We believe that success of foundation models to enhance downstream algorithmic fairness during adaptation phases by allowing changes to better reflect marginalized communities, that have been documented by existing



(a) (b) (c) (d) (e)

Figure 1. Simplified example of data with a mixture of heterogeneous distributions. The aggregated distribution, commonly assumed for training machine learning models, is represented in Figure (a) by a black dotted line D_m, while the underlying distributions are depicted in blue D_1, orange D_2, green D_3, and red D_4. In this example, all distributions are assumed to be Gaussian with equal weight, which is not reflective of real-world scenarios where marginalized communities often have smaller data sizes. Despite this simplification, the aggregated distribution still differs significantly from all the underlying distributions. Learning based solely on the aggregated distribution not only fails to accurately represent any of the underlying distributions, as shown in (a), but also risks missing variations if the dimensions of the underlying distributions differ, as shown in (b-e). While simply reweighting or adding more data points to marginalized distributions is not helpful, the complexities of the distributions can significantly impact the learning process and should be reflected in the method of aggregation.

work (Mao et al., 2023), are due to better arrangements of data at such smaller scale. However, accurately grouping data requires a comprehensive understanding of underlying distributions.

The challenge of training machine learning models under heterogeneous data distributions is a significant area of research, particularly within the domain of federated learning (FL). In FL, models are trained collaboratively across clients that possess diverse datasets, reflecting real-world scenarios where data is not uniformly distributed (McMahan et al., 2017). While conventional FL treats each client as a unique distribution, there are often underlying sub-distributions within the broader heterogeneous dataset. To address this, advanced clustering techniques has been employed to identify and group these distinct sub-distributions, enabling more targeted model training (Ghosh et al., 2019; Malekmohammadi et al., 2024).

An additional challenge lies in distinguishing low-quality data distributions from those representing marginalized communities. There is no universal definition of low-quality data; however, techniques designed to mitigate data poisoning and adversarial attacks can inadvertently misclassify data from marginalized communities as low-quality. While channeling inferences based on data distribution may help address poisoning and adversarial attacks, extensive research is needed to differentiate between truly low-quality data and unique characteristics of marginalized groups' data distributions.

Call to Action: Metrics for Representation Disparity via Manifold Embedding: Our second call to action, is a call to address the potential flattening of latent dimensions learned during training for marginalized communities. The flattening of latent dimensions suggests that the model may not fully capture the complexities of their distributions, unlike those of majority groups.

To better grasp this, consider Figure 1. If distribution D_4 (sub-figure (e)) relies primarily on dimensions x_3 and x_5, learning dimensions x_1, x_2, x_3, x_4 from other distributions might obscure crucial nuances of x_5. While distributions can share common elements such as sentence structure, alphabet, or cultural norms in language, they also possess unique dimensions, e.g., consider x_3 in Figure 1 which is a shared dimension between distributions D_2 and D_4. Dedicating model capacity specifically to learn these distinct dimensions would enable far more accurate representation.

Based on the theoretical foundation of manifold embeddings (Melas-Kyriazi, 2020), we need to build fair representation for marginalized communities within data (Wan, 2021). Comprehensive metrics to measure the multifaceted impact of data selection bias on model behavior are lacking, hindering effective evaluation. Existing fairness benchmarks often lack robustness, present wide range of ambiguities, social science pitfalls (Blodgett et al., 2021; Gallegos et al., 2023) and fail to address core issues within learned representations. Also, there is need for metrics that go beyond measuring disparities in downstream tasks. In single task machine learning models, disparities are typically measured within the output space specific to that particular downstream task. Consequently, most available metrics that consider fairness and disparities among different groups focus on evaluations within this output space. Foundation models, however, diverge from this paradigm as they are intended to serve as versatile representations applicable to a wide array of tasks, including those that may emerge in the future. Given the expansive scope of application for foundation models, it becomes imperative to consider metrics that are robust enough to assess group disparities within the representation space. Evaluating the learned manifold involves assessing various geometric properties that can capture richness and nuances of the data representation. We want to emphasize that simple distance metrics in embedding spaces, as proposed in the literature (Zhao et al., 2017),

fail to fully assess the quality of embeddings for different communities. Collapsing many dimensions during the embedding process leads to a loss of nuance for marginalized communities, potentially assigning identical embeddings to distinct concepts. Lipschitz continuity can be used to evaluate how smoothly the manifold transitions between different regions.Szegedy et al. (2013) showed that for CNNs, a lower Lipschitz constant indicates more robust features and results in higher generalizability.

Call to Action: Guiding Capacity via Mixture of Expert Models: Our next call to action lies in developing measures that both quantify differences between distributions within the embedding space and determine the capacity a model should allocate to each distribution.

Mixture of Experts (MoE) models offer a promising approach to address the challenges inherent in training with heterogeneous data (Fedus et al., 2022). Their architecture, consisting of specialized "expert" neural networks and a learned "gating" network, enables intelligent routing of inputs to the most appropriate expert(s). This allows for targeted training on specific data aspects or subdistributions, making MoEs well-suited to the nuanced complexities we've outlined.

Furthermore, recent research into sparse networks suggests that a substantial portion of model capacity may be dispensable without sacrificing accuracy on majority-class data. This opens up potential avenues for exploration using techniques like Parameter-Efficient Tuning (PEFT) (Gordon et al., 2023) or LoRA (Hu et al., 2021) to optimize MoE performance and reduce computational overhead.

In addition, PEFT has a potential to facilitate collaborative federated model training with low-resource marginalized communities. By tailoring resource allocation based on the size of individual communities distribution, we can ensure privacy protection in federated paradigms while avoiding the mismatched noise levels that often arise in differential privacy applications sensitive to distribution size (see Section 3).

Call for Action: Model-guided Data Collection: Our last call for action is to address a fundamental challenge faced by marginalized communities which is the non-uniform sampling of their data (see Section 5). This leads to frequent distribution shifts (as discussed in Section 3) and hinders model generalization. To counter this, we propose automated methods to identify sparse areas within marginalized data distributions and strategically employ active learning techniques (Wang et al., 2017). Active learning enables models to proactively query for the most informative data points, guiding targeted curation and collection efforts to improve model performance on underrepresented areas (Holzinger et al., 2016).

Recent fine-tuning techniques like RLHF and RLAIF (see Section 5) can be adapted to guide active data creation mechanisms, too. This offers a targeted approach to alleviate data sparsity within marginalized communities (Hemmat et al., 2023). Our proposal emphasizes strategic data acquisition over simplistic model scaling or unguided data curation. While existing model capacity might be sufficient to learn marginalized data distributions, focused data collection is crucial. However, it's important to note that solely relying on model-identified sparsity may not fully reflect real-world data complexities. This approach primarily serves to improve the efficiency of data collection efforts. We fully acknowledge the potential for model-guided data collection to perpetuate existing biases. We emphasize that our approach is not a replacement for addressing fundamental data quality issues like the lack of digitized data in low-resource languages or marginalized communities. However, we believe that uncritically collecting more data can also reinforce existing disparities. Our method intentionally seeks out underrepresented areas, aiming to break those cycles. Our goal in this call for action is to optimize the data collection process within existing constraints. By identifying areas where the model's representations are inadequate, we can target data collection to fill gaps and broaden the model's understanding of underrepresented distributions instead of blindly increasing the dataset size. The inspiration from active learning highlights this point – the model itself suggests where to focus collection efforts. We propose a dynamic process where the model actively identifies its representational shortcomings. This targeted approach could potentially be more effective in broadening a model's understanding, especially in multilingual settings where conceptual gaps are readily detectable e.g., the significant variation, or even near-zero distance, between representations of distinct concepts across or within languages.

Finally, we believe covering all the nuances of the representation disparity and how all other courses of disparity stemmed from it by studying it thought the ML pipeline is an important aspect to picture and show the complexity of the cascading phenomena. Our focus on addressing this root cause through training procedures and novel metrics is an intentional and focused call to action. Current metrics are insufficient, and we believe developing new evaluations and metrics targeting capacity and embedding gaps is crucial for tackling representation disparities. Furthermore, as discussed earlier, we believe overemphasizing on solutions such as scaling laws and indiscriminate data collection can mask biases favoring majority groups. Recent literature on pruning indicates that we may be using unnecessarily large models for majority representation and it is even an *overkill". In this position paper, we advocate for optimizing model capacity during training to ensure inclusivity across diverse distributions, i.e., MoE models. We emphasize that our calls to action are intentionally interconnected, forming a unified framework to address the cascading effect and prioritizing one call to action over another would miss the systemic nature of the problem.

Acknowledgments

We would like to thank Hugo Larochelle for insightful discussions at early stages of the project, and Stephen Pfohl for his valuable feedback on later drafts. We also thank our reviewers for their insightful comments and constructive feedback.

Impact Statement

Addressing the multifaceted problem of disparities in foundation models requires a holistic sociotechnical approach that goes beyond isolated technical fixes. While in previous section, we solely focused on technical gaps and position our work around the cascading disparity towards better training paradigms, we would like to emphasize on the need for **community engagement** and involve marginalized communities throughout the development process to ensure that their perspectives and concerns are addressed.

Foundation models democratize AI by empowering smaller organizations and communities with limited data or expertise. They unlock previously impossible applications like creative text generation and code translation. However, they also introduce unique challenges that emerge after deployment which impact marginalized community. While a comprehensive discussion of the societal impacts of foundation models is beyond the scope of this work, to conclude our paper, next, we highlight three critical areas that directly affect marginalized communities and require urgent attention.

Measurements beyond Representation Disparity: The size and complexity of foundation models make it difficult to comprehensively assess their disparities. Traditional disparity detection methods often fall short due to the model's ability to mask disparities in subtle ways. Foundation models, specifically multi-modal versions, often process multiple types of data, such as text, images, and audio. This means that disparities can manifest in different forms, making it challenging to develop a universal evaluation framework. Additionally, disparities can emerge in different contexts and evolve over time. Static evaluation methods may not capture these dynamic patterns. Finally, currently, no widely accepted set of metrics exists to measure disparities in foundation models. Different researchers and organizations may prioritize different fairness criteria, which makes comparisons and bench marking very difficult and challenging.

Monopoly Effects: Unlike traditional ML models, training foundation models necessitates vast amounts of data and specialized hardware, often inaccessible to most organizations and researchers. This creates a high barrier to entry for marginalized communities, and leads to a concentration of power among a few large tech companies with the resources to develop and control these models. This concentration can stifle competition and innovation, as smaller players are often unable to challenge the dominance of large players with their proprietary foundation models. Moreover, as more organizations rely on foundation models from a few providers to power their products and services, a widespread dependence on these models emerges which lead to propagation of disparities in various downstream tasks that are stemmed from a single model. This can also create a lock-in effect, which makes it difficult to transition to alternative models or providers, as entire ecosystems become built upon a limited set of foundation models. This concentration of power grants a few companies significant control over the development and direction of AI research and applications while the root of disparities are the same and create a systematic way that disparities impact society, specifically, marginalized communities.

Long-term Impact: Foundation models undergo training using data obtained from internet scraping. As mentioned in previous sections, this data is not selectively chosen; rather, it is randomly sampled from the internet. Consequently, the distribution of this data is biased, reflecting the cultures and attributes of regions with greater internet access and usage. Given that foundation models find application in numerous generative tasks across various modalities (e.g., text, image, video, music, etc.), this exacerbates the existing gap in the representations of data on the internet. The data used to train these foundation models in subsequent iterations contributes to a snowball effect, potentially resulting in almost zero representations of certain marginalized groups.

Evaluation for downstream task: Addressing various biases towards marginalized communities in foundation models requires an in-depth understanding of interconnected harms and the development of holistic solutions. Future research should focus on the development of robust metrics and approaches that surpass surface-level evaluations. Existing work suggests conflicting empirical results about the relations of intrinsic biases and extrinsic biases (Gupta et al., 2022; Kiela et al., 2022; Stafanovičs et al., 2022; Mohanty et al., 2022). We encourage future research to focus on theoretical foundations to analyze the relations between various intrinsic biases and extrinsic biases through the lens of representation disparity.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Ali, J., Kleindessner, M., Wenzel, F., Budhathoki, K., Cevher, V., and Russell, C. Evaluating the fairness of discriminative foundation models in computer vision. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 809–833, 2023.

Allman, D. The sociology of social inclusion. *Sage Open*, 3(1):2158244012471957, 2013.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Benkler, N., Mosaphir, D., Friedman, S., Smart, A., and Schmer-Galunder, S. Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075*, 2023.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Black, E., Raghavan, M., and Barocas, S. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, 2022.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, 2021.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Carlini, N. and Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint arXiv:1801.01944*, 2018.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022a.

Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., and Tramer, F. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022b.

Cohen, J. P., Luck, M., and Honari, S. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 529–536. Springer, 2018.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., and Narasimhan, K. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Du, M., Yang, F., Zou, N., and Hu, X. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020.

Dwivedi, S., Ghosh, S., and Dwivedi, S. Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4), 2023.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pp. 119–133. PMLR, 2018.

Dziri, N., Milton, S., Yu, M., Zaiane, O., and Reddy, S. On the origin of hallucinations in conversational models: Is it the datasets or the models? *arXiv preprint arXiv:2204.07931*, 2022.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Ganesh, P. An empirical investigation into benchmarking model multiplicity for trustworthy machine learning: A case study on image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4488–4497, 2024.

Ganesh, P., Chang, H., Strobel, M., and Shokri, R. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1789–1800, 2023.

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

Ghosh, A., Hong, J., Yin, D., and Ramchandran, K. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

Gorban, A. N. and Tyukin, I. Y. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2118):20170237, 2018.

Gordon, A., Dujmovic, J., Izacard, G., Riedel, S., and Srinivasan, K. Morphing transformers for fine-tuning massive language models with minimal effort. *arXiv preprint arXiv:2302.03102*, 2023.

Guo, C., Rana, M., Cisse, M., and Robby. Deep counterfeit: A black-box attack against autonomous driving models. *arXiv preprint arXiv:1801.10578*, 2018.

Gupta, P., Kassner, N., and Schütze, H. How gender debiasing affects internal model representations, and why it matters. *arXiv preprint arXiv:2204.06827*, 2022. URL https://arxiv.org/abs/2204.06827.

Hall, M., Gustafson, L., Adcock, A., Misra, I., and Ross, C. Vision-language models performing zero-shot tasks exhibit disparities between gender groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2778–2785, 2023.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Hargittai, E. Minding the digital gap: Why understanding digital inequality matters. In *Media perspectives for the 21st century*, pp. 231–240. Routledge, 2011.

Hashemizadeh, M., Ramirez, J., Sukumaran, R., Farnadi, G., Lacoste-Julien, S., and Gallego-Posada, J. Balancing act: Constraining disparate impact in sparse models. *arXiv preprint arXiv:2310.20673*, 2023.

Hemmat, R. A., Pezeshki, M., Bordes, F., Drozdzal, M., and Romero-Soriano, A. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023.

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., and Palade, V. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2): 119–131, 2016.

Hooker, S. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4), 2021.

Hooker, S., Courville, A., Clark, G., Dauphin, Y., and Frome, A. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Hu, E. J., Shen, Y., Wallis, P., Neubig, G., Riloff, E., and Brunk, M. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jha, A., Prabhakaran, V., Denton, R., Laszlo, S., Dave, S., Qadri, R., Reddy, C. K., and Dev, S. Beyond the surface: A global-scale analysis of visual stereotypes in text-to-image generation. *arXiv preprint arXiv:2401.06310*, 2024.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jiang, Z., Zhang, C., Talwar, K., and Mozer, M. C. Characterizing structural regularities of labeled data in overparameterized models. *arXiv preprint arXiv:2002.03206*, 2020.

Kiela, D., Foka, P., Aktas, O. S., Lin, M. Y.-J., Baudia, P., and Celikyilmaz, A. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. In *15th International Conference on Language Resources and Evaluation*, 2022.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pp. 6565–6576. PMLR, 2021.

Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Ma, X., Wang, Z., and Liu, W. On the tradeoff between robustness and fairness. *Advances in Neural Information Processing Systems*, 35:26230–26241, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Maene, J., Li, M., and Moens, M.-F. Towards understanding iterative magnitude pruning: Why lottery tickets win. *arXiv preprint arXiv:2106.06955*, 2021.

Malekmohammadi, S., Taik, A., and Farnadi, G. Mitigating disparate impact of differential privacy in federated learning through robust clustering. *arXiv preprint arXiv:2405.19272*, 2024.

Mao, Y., Deng, Z., Yao, H., Ye, T., Kawaguchi, K., and Zou, J. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Melas-Kyriazi, L. The mathematical foundations of manifold learning. *arXiv preprint arXiv:2011.01307*, 2020.

Mohanty, S., Serdyukov, P., Petrov, D., and Nadeem, M. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2204.01717*, 2022.

Molamohammadi, M., Taïk, A., Le Roux, N., and Farnadi, G. Unraveling the interconnected axes of heterogeneity in machine learning for democratic and inclusive advancements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–12, 2023.

Moradi, A. and Farnadi, G. Tidying up the conversational recommender systems' biases. *arXiv preprint arXiv:2309.02550*, 2023.

Nadeem, M., Bethke, A., and Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

Natekar, P. and Sharma, M. Representation based complexity measures for predicting generalization in deep learning. *arXiv preprint arXiv:2012.02775*, 2020.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019.

Passi, S. and Barocas, S. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 39–48, 2019.

Pedreschi, D., Ruggieri, S., and Turini, F. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*, pp. 581–592. SIAM, 2009.

Prates, M. O., Avelar, P. H., and Lamb, L. C. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381, 2020.

Qadri, R., Shelby, R., Bennett, C. L., and Denton, E. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 506–517, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rao, A., Khandelwal, A., Tanmay, K., Agarwal, U., and Choudhury, M. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*, 2023.

Ravi, S. and Beatson, A. Amortized bayesian meta-learning. In *International Conference on Learning Representations*, 2018.

Richardson, R., Schultz, J. M., and Crawford, K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, 94:15, 2019.

Samory, M., Sen, I., Kohne, J., Flöck, F., and Wagner, C. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pp. 573–584, 2021.

Scherrer, N., Shi, C., Feder, A., and Blei, D. M. Evaluating the moral beliefs encoded in llms. *arXiv preprint arXiv:2307.14324*, 2023.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., and Lockhart, J. W. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171, 2020.

Sen, I., Samory, M., Wagner, C., and Augenstein, I. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. *arXiv preprint arXiv:2205.04238*, 2022.

Stafanovičs, T., Grundkiewicz, R., and Schütze, H. Mitigating gender bias in neural machine translation with target-side monolingual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1798–1811, 2022.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belgrave, E., Abebe, G., Frauenholz, S., et al. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Suresh, H. and Guttag, J. V. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8), 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tatman, R. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 53–57, 2017.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction apis. *arXiv preprint arXiv:1609.02943*, 2016.

Tran, C., Fioretto, F., Kim, J.-E., and Naidu, R. Pruning has a disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 35:17652–17664, 2022.

Wan, A. Fairness in representation for multilingual nlp: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*, 2021.

Wang, C. and Sennrich, R. On exposure bias, hallucination and domain shift in neural machine translation. *arXiv preprint arXiv:2005.03642*, 2020.

Wang, K., Chen, G. S., Diao, R., Wai, A. P., Korosoglou, G., Cheng, L.-F., Chen, Z., and Sethi, I. K. Active learning with neural networks for personalized medicine. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 135–142. IEEE, 2017.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Williams, C. C. and White, R. Conceptualising social inclusion: some lessons for action. In *Proceedings of the Institution of Civil Engineers-Municipal Engineer*, volume 156, pp. 91–95. Thomas Telford Ltd, 2003.

Yang, X., Li, Y., Zhang, X., Chen, H., and Cheng, W. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*, 2023.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1497–1510, 2017.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending against neural fake news. In *Advances in neural information processing systems*, pp. 9051–9062, 2019.

Zhang, R., Liao, S., Li, Z., Lei, Z., and Jain, A. K. Mitigating bias in face recognition: The impact of demographic-balanced datasets. *arXiv preprint arXiv:1811.00483*, 2018.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.