Journal of Data-centric Machine Learning Research (2024) Submitted 5/24; Revised 8/24; Published 9/24

On Catastrophic Inheritance of Large Foundation Models

Hao Chen HAOC3@ANDREW.CMU.EDU
Carnegie Mellon University

Bhiksha Raj BHIKSHAR@ANDREW.CMU.EDU
Carnegie Mellon University

Xing Xie XINGX@MICROSOFT.COM
Microsoft Research

Jindong Wang JINDONG.WANG@MICROSOFT.COM
Microsoft Research, William & Mary

Reviewed on OpenReview: *https://openreview.net/forum?id=fONiOrnjLE&referrer=%5BAuthor%20Console%5D*

Editor: Andreas Kirsch

Abstract

Large foundation models (LFMs) are claiming incredible performances. Yet great concerns have been raised about their mythic and uninterpreted potentials not only in machine learning, but also in various other disciplines. In this position paper, we propose to identify a neglected issue deeply rooted in LFMs: *Catastrophic Inheritance*, describing the weaknesses and limitations inherited from biased large-scale pre-training data to behaviors of LFMs on the downstream tasks, including samples that are corrupted, long-tailed, noisy, out-of-distributed, to name a few. Such inheritance can potentially cause catastrophes to downstream applications, such as bias, lack of generalization, deteriorated performance, security vulnerability, privacy leakage, and value misalignment. We discuss the challenges behind this issue and propose "UIM", a framework to *Understand* the catastrophic inheritance of LFMs from both pre-training and downstream adaptation, *Interpret* the implications of catastrophic inheritance on downstream tasks, and how to *Mitigate* it. UIM aims to unite both the machine learning and social sciences communities for more responsible and promising AI development and deployment.

Keywords: Catastrophic Inheritance, Pre-training Data, Foundation Models

1 Introduction

In the rapidly evolving landscape of machine learning, large foundation models (LFMs), such as CLIP (Radford et al., 2021; Cherti et al., 2023), GPT (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023), PaLM-2 (Anil et al., 2023), LLaMA (Touvron et al., 2023a,b), Stable Diffusion (Rombach et al., 2022), Gemini (Google, 2023), Time-LLM (Jin et al., 2023), etc, have emerged as a cornerstone (Bommasani et al., 2021) for numerous real-world

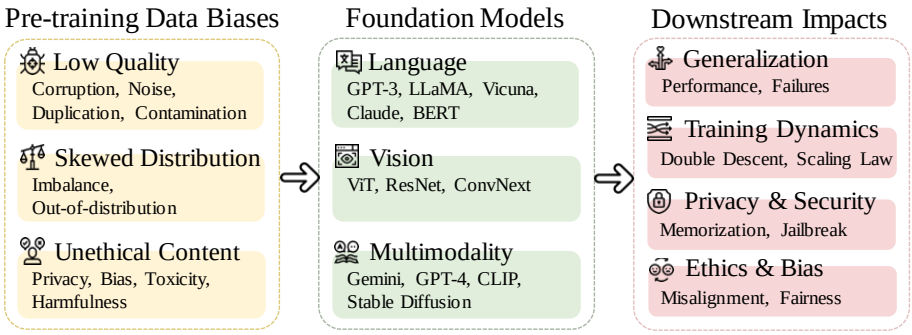

Figure 1: Illustration of catastrophic inheritance. Large foundation models pre-trained on biased datasets may cause significantly malicious consequence to various downstream tasks (rf. Table 1).

tasks. Characterized by their large parameter sizes and extensive training on large-scale data (Sevilla et al., 2022), LFMs have demonstrated remarkable abilities such as zero-shot learning (Radford et al., 2021; Gruver et al., 2024) and in-context learning (Radford et al., 2019; Brown et al., 2020; Gao et al., 2020b; Kaplan et al., 2020; Zhao et al., 2021; Wei et al., 2022; Olsson et al., 2022; Rasul et al., 2023), and impressive transfer performance across various tasks (Zhuang et al., 2020; He et al., 2021; Nakkiran et al., 2019; Kaplan et al., 2020). As LFMs claim promising performances in almost every discipline from computer science, natural science, to social science, it is urgent yet challenging to fully evaluate and understand their capabilities, limitations, and failures.

This paper proposes a phenomenon and novel research direction – *Catastrophic Inheritance*, describing that LFMs pre-trained on increasingly large-scale but biased datasets can cause potentially significant and catastrophic consequences to downstream tasks (Caballero et al., 2022; Schaeffer et al., 2023). As evidenced in Table 1, various user applications that rely on LFMs are affected by potentially biased pre-training datasets from multiple aspects, including ethics (Forbes, 2023; Thiel, 2023), security (Wang et al., 2018; Zhang et al., 2022), generalization (Chen et al., 2024b), language understanding (Jin et al., 2024), and culture bias (Boston.com, 2023), to name a few. For example, LAION-5B (Schuhmann et al., 2022), the popular pre-training dataset for Stable Diffusion and many other LFMs, is reported to contain harmful content (Birhane et al., 2023), such as child sexual abuse material (Forbes, 2023), which was then inherited to Stable Diffusion models to generate similar harmful contents. Existing research also shows that biases in the pre-training data inevitably perturb and maliciously affect the generalization and behaviors of LFMs (Dodge et al., 2021; Chen et al., 2024b; Dong et al., 2023; Longpre et al., 2023). Perhaps more alarmingly, the detrimental effects of biases might be concealed superficially after fine-tuning on specific downstream tasks (Jain et al., 2023b; Qi et al., 2023), which may consequently raise safety and security concerns in the deployment (Carlini et al., 2023a; Gu et al., 2023; Mallen et al., 2022). In essence, despite the difference in architectures and proxy pre-training tasks of LFMs (Vaswani et al., 2017; Ronneberger et al., 2015), the myth of their training behaviors and capabilities (Nakkiran et al., 2019; Power et al., 2022; Kaplan et al., 2020) largely inherit from the opaque and large-scale pre-training datasets (Entezari et al., 2023; Elazar et al., 2023).

With the rapid evolution of LFMs, scaling the dataset from web-collected contents (Raffel et al., 2020; Gao et al., 2020a; Schuhmann et al., 2022; Byeon et al., 2022; Computer, 2023; Penedo et al., 2023; Soldaini et al., 2023) becomes a convention to improve model generalization, which avoids heavy human efforts of curation and annotation (Birhane et al., 2023). However, does scaling really beat (the effect of) biases? The increasingly scaled Internet data also inevitably contains more *imbalanced* (Reed, 2001; Zhu et al., 2023a; Parashar et al., 2024), *duplicated* (Lee et al., 2022; Hernandez et al., 2022; Tirumala et al., 2023; Elazar et al., 2023; Xu et al., 2023b), *corrupted* (Luccioni and Viviano, 2021; Birhane et al., 2023; Carlini et al., 2023a; Yang et al., 2023b), *contaminated* (Marone and Van Durme, 2023; Wei et al., 2023b; Oren et al., 2023; Deng et al., 2023; Jiang et al., 2024), *noisy* (Kolesnikov et al., 2020; Schuhmann et al., 2022; Chen et al., 2024b), and even *unethical* and *biased* (Forbes, 2023; Thiel, 2023; Jin et al., 2024) samples. Even more recently, synthetic data has been widely involved in the pre-training of large language models (Gunasekar et al., 2023) and time-series foundation models (Dooley et al., 2024). While showing promising downstream performance, the limitation of synthetic data should also be considered (Alemohammad et al., 2023). Furthermore, some of the advanced and proprietary LFMs such as GPT-4 (OpenAI, 2023) and Gemini (Google, 2023) do not open-source their training data. The huge volume, high complexity, and black-box nature of the pre-training data make it economically expensive and technically impossible to detect and remove all the biased samples, which thus maliciously affect the LFMs' behavior and generalization.

In this paper, we propose *UIM*, a general research framework to understand, interpret, and mitigate the catastrophic inheritance of LFMs to the downstream tasks. Despite the prosperous development and research to improve the generalization of LFMs, addressing catastrophic inheritance has received limited attention and presents a few unsolved challenges. First, it remains unclear how the pre-training data biases will directly affect the generalization properties and the training dynamics (Nakkiran et al., 2019; Kaplan et al., 2020; Power et al., 2022) of these models at the pre-training stage, which may inherit to the subsequent tasks (Caballero et al., 2022; Dar et al., 2021). Second, it remains unknown how to interpret the effects of biased pre-training data on downstream tasks and the fundamental reasons for these effects from LFMs (Bender et al., 2021; Jain et al., 2023b; Chen et al., 2024b). The lack of comprehensive evaluation and proper metrics of LFMs beyond the performance of downstream tasks is one of the most essential reasons impeding our understanding and interpretation of catastrophic inheritance (Sun et al., 2023; Lee et al., 2023; Schaeffer et al., 2023; Tong et al., 2024). Third, due to the black-box and complex nature of LFMs and pre-training datasets, it becomes notoriously difficult to mitigate the malicious effects of pre-training biases on downstream tasks (Oren et al., 2023; Chen and Yang, 2023; Chen et al., 2024b; Zhang et al., 2024b), without re-train the model from scratch. To overcome these challenges, UIM involves three aspects:

- **Understanding** the catastrophic inheritance from pre-training dynamics, generalization behaviors, scaling laws, effects on downstream tasks, with more comprehensive evaluation benchmarks and effective metrics of LFMs.
- **Interpreting** the fundamental sources in LFMs that lead to catastrophic inheritance on generalization to downstream tasks, both empirically and theoretically.

Table 1: Realistic examples of catastrophic inheritance from published papers or news.

Example	Domain	Source
Stable Diffusion models was trained on Laion-5B, which contains hundreds of harmful images of child sexual abuse material (CSAM). Then, the model was reported to memorize during training and generate CSAM at production.	Ethics and privacy	(Birhane et al., 2023; Forbes, 2023; Thiel, 2023)
At least 50% of poisoning, adversarial, and backdoor vulnerabilities will be inherited from pre-training data to fine-tuned models, which can be easily triggered at the deployment. Jailbreaks may also relate to pre-training biases.	Security	(Wang et al., 2018; Zhang et al., 2022; Carlini et al., 2023a; Zou et al., 2023)
An MIT student asked AI to make her headshot more 'professional.' It gave her lighter skin and blue eyes. Country bias also found in language models.	Bias	(Boston.com, 2023; Wang et al., 2023c)
Fine-tuning LLMs on only 10 adversarially designed or even benign samples leads to degradation of safety alignment, which costs less than $0.2 using API.	Misalignment	(Qi et al., 2023)
Noisy labels contained in pre-trained data always hurt downstream OOD performance; more than 10% noisy data will hurt in-domain performance.	Generalization	(Chen et al., 2024b)
Large language models like GPT-3.5 exhibited an accuracy reduction of 18.12% when answering non-English medical questions. Similar for coding tasks.	Model behaviors	(Jin et al., 2024; Zheng et al.)
Noise in the pre-training data strengthen the double descent phenomena, where the critical point of LFMs overfitting/memorizing data appears earlier.	Training dynamics	(Nakkiran et al., 2019)

- **Mitigating** catastrophic inheritance on downstream tasks in (partially) black-box paradigms without re-training LFMs from scratch, access to full architecture/weights of LFMs, and access to large-scale pre-training datasets.

UIM stands out as a set of under-explored research directions that could trigger many new opportunities not only in connecting traditional machine learning efforts to LFMs but also in unprecedented interpretation of LFMs, including vision, language, and most importantly, social sciences. Since the impacts of LFMs lie not only in the algorithmic level, but in societal level that matters to everyone. The involvement of social sciences is indispensable to help researchers better evaluate the capabilities of models, measure societal impact, design human study, and delve into all aspects of society for risk management. We hope these research topics will facilitate a better understanding on the generalization of foundation models from both the stage of pre-training and downstream tasks transferring, which ultimately helps us curate more high-quality pre-training datasets and build more promising models. The remainder of the paper is organized as follows. In Section 2, we introduce the relevant background, formal definition of catastrophic inheritance, and preliminary studies in the relevant fields. We then identify the challenges of understanding and solving catastrophic inheritance in Section 3 and present our proposals for future research in each dimension with more details in Section 4. At the end, we conclude this position paper in Section 5.

2 Catastrophic Inheritance

In this section, we present a comprehensive review of literature related to catastrophic inheritance, defined as:

Definition 1 *Catastrophic Inheritance (CI) refers to as the catastrophic and malicious impacts of adapting large foundation models $\mathcal{M}$ on downstream tasks with data $\mathcal{D}_{\text{down}}$ and algorithm $\mathcal{A}_{down}$, which are learned and inherited from the large-scale but potentially biased pre-training data $\mathcal{D}_{\text{up}}$ with the pre-training proxy algorithm $\mathcal{A}_{\text{up}}$:*

$$\text{CI} = g\left(\mathcal{D}_{\text{down}}, f\left(\mathcal{D}_{\text{up}}, \mathcal{M}, \mathcal{A}_{\text{up}}\right), \mathcal{A}_{\text{down}}\right), \tag{1}$$

where f corresponds to the pre-trained model that encompasses the change of models' behaviors, capacities, and generalization, and g models the malicious impacts on downstream subjected to both pre-training and downstream.

The catastrophic inheritance thus models a function of both downstream and pre-training model, dataset, and algorithm. We review the recent realistic examples of catastrophic inheritance (Section 2.1), the related works from the biases in pre-training data $\mathcal{D}_{\text{up}}$ (Section 2.2), the potential impacts of such biases g on downstream tasks (Section 2.3), and the rather underexplored mitigation strategies on the malicious effects of them, which reduces g, (Section 2.4), especially the black-box methods due to the limited access of LFMs.

2.1 Realistic Examples of Catastrophic Inheritance

Here, we present realistic examples that underscore the concept of catastrophic inheritance, as shown in Table 1.

Studies (Wang et al., 2018; Rezaei and Liu, 2020) have indicated that fine-tuned models can inherit issues from pre-trained models containing backdoor vulnerabilities. This risk is amplified in large-scale pre-training datasets, as Carlini et al. (2023a) demonstrated, which are prone to being poisoned, thus adversely affecting downstream tasks. Zhang et al. (2022) found a significant probability (approximately 50%) that downstream fine-tuned models inherit adversarial and backdoor vulnerabilities from their pre-trained counterparts. The massive capacity of LFMs often leads to the memorization of these harmful samples, which could manifest at deployment, thereby raising severe security, privacy, and bias concerns (Birhane et al., 2023; Qi et al., 2023).

Moreover, Chen et al. (2024b) have shown that the presence of noisy labels in pre-training data consistently undermines performance in out-of-distribution tasks. The noise inherent in pre-training data also affects the training dynamics of LFMs (Nakkiran et al., 2019). Jin et al. (2024) reported a notable decrease in accuracy (about 18.12%) by GPT-3.5 in responding to non-English medical inquiries. The corresponding findings were reported by Zheng et al., showing that language models pre-trained in English tend to outperform those trained in Chinese on Chinese-language tasks. However, comprehensive methods for understanding, interpreting, and mitigating the effects of catastrophic inheritance in LFMs remain largely undeveloped.

Table 2: Common pre-training data biases identified in previous literature.

Bias Type	Biased Data & Def.	Malicious Effects	Source
Low Quality	Duplication: Exactly the same and semantically similar content/samples	Memorization, privacy risks	Elazar et al. (2023); Carlini et al. (2022); Hernandez et al. (2022)
Low Quality	Corruption/Noise: Unnatural and Unmatched inputs and supervision	Deteriorated generalization and performance on downstream	Elazar et al. (2023); Fan et al. (2023a); Kreutzer et al. (2022)
Low Quality	Contamination: Leakage of testing samples to training data	Broken and inaccurate evaluation	Roberts et al. (2023); Schaeffer (2023); Jiang et al. (2024)
Skewed Dist.	Imbalance: Concepts clusters form different and imbalanced proportion	Biased predictions from rare concepts with worse performance	Xu et al. (2023c); Zhu et al. (2023a); Parashar et al. (2024)
Unethical Content	Biases, toxicity, and harmfulness	Harmful generation	Zou et al. (2023); Sun et al. (2024)

2.2 Biases in the Pre-training Data

We discuss the common biases in the pre-training data $\mathcal{D}_{\text{up}}$ identified in the previous literature, as shown in Table 2. We summarize three types of pre-training data biases: low quality, skewed distribution, and unethical content.

Low quality. Low-quality training samples can be prevalent in large-scale, web-crawled pre-training datasets, which includes and not limits to data duplication, corrupted, noisy, and contaminated data. These biases directly affect LFMs' behaviors and capabilities on various downstream tasks (Hall et al., 2022).

Repeated samples have been reported as a common occurrence in the pre-training data. Studies (Kandpal et al., 2022; Elazar et al., 2023) have revealed a significant percentage of duplicates in datasets such as RedPajama (Computer, 2023) and Laion-2B-en (Schuhmann et al., 2022). This repetition not only affects the efficiency of the learning process but also presents memorization issues that lead to privacy risks, as discussed by Carlini et al. (2022) and Lee et al. (2022). Hernandez et al. (2022) also studied the scaling laws and interpretability of training language models duplicated in a systematic manner, confirming that repeated data can negatively impact the learned structures crucial for generalization. Many recent research has focused on the de-duplication of the pre-training data with various techniques (Coupette et al., 2021; Abbas et al., 2023) and found improved generalization when training on filtered and deduplicated data (Penedo et al., 2023; Tirumala et al., 2023).

Corrupted and noisy samples and supervision are prevalent, encompassing broader issues than traditional noisy label learning (Natarajan et al., 2013), including unmatched pairs in multimodal datasets and low-quality elements in self-supervised pre-training. Kreutzer et al. (2022) highlighted the presence of such low-quality texts in web-scale datasets, especially in low-resource languages. Similarly, Gunasekar et al. (2023) and Zheng et al. observed performance disparities in language models trained on different language codes and sources. Jain et al. (2023a) supported the idea that structured data leads to better results in tasks

such as code generation. Recent trends include using synthetic data for pre-training, which, if of low quality, can also introduce the corruption to pre-training. Noise and corruption in the pre-training data can impose impacts on the behaviors of the models and generalization to downstream tasks in various dimensions (Longpre et al., 2023; Chen et al., 2024b).

Data contamination in LFMs, where training data overlaps with test data, has also been increasingly recognized as problematic. It challenges our understanding of LFMs' true capabilities (Dodge et al., 2021; Yang et al., 2023a; Roberts et al., 2023; Li, 2023b; Deng et al., 2023; Jiang et al., 2024). Studies Schaeffer (2023) showed that training on test data can disrupt expected scaling laws and induce grokking behaviors. Li and Flanigan (2023) found that LLMs perform differently depending on the date of creation of the test data. Recognizing the overlap of training and testing data, new metrics for detecting contamination have emerged, such as loss difference (Wei et al., 2023b), model-based (Yang et al., 2023a), perplexity (Li, 2023a), and black-box method (Oren et al., 2023) without access to pre-training data or model.

Skewed Distribution. The concepts/clusters/subsets in the web-collected pre-training data often exhibit long-tailed distributions (Reed, 2001) that are difficult to re-balance at scale, casting challenges to most of the self-supervised LFMs (Kandpal et al., 2023). The imbalance skews LFMs' capabilities towards more frequent concepts, as demonstrated by Zhu et al. (2023a). For instance, the CLIP (Radford et al., 2021) and MetaCLIP (Xu et al., 2023c) models show better generalization than those trained on LAION-400M (Schuhmann et al., 2021; Cherti et al., 2023), due to their "balanced" data curation strategy. Instead of naively scraping web data, CLIP and MetaCLIP collected at most 20K image-text pairs for each of 500K visual concepts. Despite efforts to data balancing, many visual concepts remain underrepresented, with fewer than 20K samples (Xu et al., 2023c). Consequently, applications reliant on pre-trained CLIP models, including vision-language chatbots (Liu et al., 2023a; OpenAI, 2023) and text-to-image generative models (Rombach et al., 2022), also fail to recognize or generate images featuring rare concepts (Parashar et al., 2024).

Unethical Content. The pre-training data for LFMs often contains content that is private, harmful, biased, or toxic, leading to significant risks in public safety, social security, and trust, particularly at the deployment of these models. Inherent biases, including gender (Kotek et al., 2023b), cultural (Tao et al., 2023), racial biases (Omiye et al., 2023), and stereotypes (Ma et al., 2023), are often reflected in these models, most likely inherited from the pre-training data. Unsafe and harmful content has been continuously reported (Jansen et al., 2022). This concern is highlighted in studies such as Dodge et al. (2021); Yao et al. (2023); Sun et al. (2024), and Kotek et al. (2023a), stressing the importance of careful data curation for model training. Addressing these issues not only improves the reliability of LFMs but also contributes to social science with more responsible AI.

Pre-training data inspection tools. Recognizing the various biases and issues in LFM pre-training data, a range of inspection tools and protocols have been developed. These include Data Portraits (Marone and Van Durme, 2023) for detecting test set leakage and model plagiarism, the Laion-2B retrieval engine (Schuhmann et al., 2022) for visualizing image-text pairs, the Text Characterization Toolkit (TCT) (Simig et al., 2022) for analyzing large dataset characteristics, searching tools (Piktus et al., 2023a,b) for qualitative analysis. The more recent Oasis (Zhou et al., 2023) offers a system for data quality assessment, and WIMBD (Elazar et al., 2023) enables fast data search and counting. The tools of more

functions documenting and understanding the pre-training data are crucial for documenting and understanding pre-training data, which is key to developing better, more effective, and well-regulated LFMs (Mitchell et al., 2022).

2.3 Potential Impacts to Downstream Tasks

In this section, we explore how biases in pre-training data potentially impact downstream tasks, i.e., g of LFM. These biases are evidenced to significantly affect training dynamics, generalization, security, and lead to misalignment and fairness issues. Understanding and interpreting these impacts is crucial for enhancing LFMs' performance and reliability, especially for applications related to society science.

Training Dynamics. Biases in pre-training data can significantly influence the training dynamics of LFMs, thus affecting their performance on downstream tasks. The double descent phenomenon (Opper, 1995; Belkin et al., 2019; Nakkiran et al., 2019), where model performance initially decreases before improving with increasing model and dataset scale, is found to be exacerbated by data biases. The transfer of double descent behavior to downstream tasks also potentially indicates a direct inheritance of affected pre-training characteristics (Dar et al., 2021). Scaling laws (Kaplan et al., 2020), which relate loss to dataset scale, model size, and training time, are critical to predicting model behaviors and the downstream performance of larger models from smaller ones. However, broken scaling laws (Caballero et al., 2022), indicating deviations in these predictions, suggest that biases in pre-training data might disrupt the expected behaviors and affect downstream generalization and performance (Cherti et al., 2023). Grokking behavior (Power et al., 2022; Varma et al., 2023a), describing the sudden spike in generalization from random to perfect levels, often occurs beyond the point of overfitting. This behavior has been linked to a transition from memorization to generalization (Kumar et al., 2023; Davies et al., 2023; Varma et al., 2023b). The correlation between training dynamics and model structure, such as induction heads in Transformers (Olsson et al., 2022; Reddy, 2023), implies that biases in pre-training data could also influence critical model functions. Understanding how these biases affect the critical data size for such dynamic changes requires further investigation (Zhu et al., 2024).

Generalization. The connection between pre-training data biases and the generalization on downstream tasks is crutial in the LFM era. Previous study (Recht et al., 2019) revealed the significant influence of data collection biases, such as those in ImageNet, on transfer performance. Although data diversity improves robustness and generalization (Fang et al., 2022; Ramanujan et al., 2023; Entezari et al., 2023), especially in real-world datasets (Fang et al., 2023; Richards et al., 2023), it often comes at the cost of quality (Nguyen et al., 2022). Merely increasing the quantity cannot guarantee the diversity always. This trade-off is exemplified in the findings of Abnar et al. (2021) and Tu et al. (2023), showing how limited data diversity and inherent biases impact the reliability and robustness of the model. The balance between intra-class and inter-class diversity also remains a complex issue to solve (Shirali and Hardt, 2023; Zhang et al., 2023a). Data pruning methods (Sorscher et al., 2022; Marion et al., 2023; Abbas et al., 2024; Fu et al., 2024) have recently been widely studied to purify the quality, improving the generalization of LLMs.

Recent studies focus more on the specific impacts of bias in pre-training data on downstream tasks, revealing nuanced effects on in-distribution and out-of-distribution performance

that may present negative correlation (Wenzel et al., 2022; Shi et al., 2023). Chen et al. (2024b) found that slight noise in pre-training data can benefit the ID performance, while always hurting the OOD performance. Hernandez et al. (2022) and Tong et al. (2024), studied data repetition and noisy image-text pairs in pre-training to increased memorization and general failures in LFMs, respectively. Yamada and Otani (2022) found that the robustified model in pre-training usually also presents robustness in downstream tasks. This evolving research area is critical to understanding catastrophic inheritance and improving the robustness and generalization of LFMs in real-world applications.

Privacy and Security. Pre-training data biases, especially those related to duplication and private information, can raise severe privacy and security issues of LFMs (Wei et al., 2023a; Bagdasaryan et al., 2023; Zou et al., 2023; Yao et al., 2023; Kumar et al., 2024; Li et al., 2024). LFMs can be elicited to verbally output private information that has been memorized at production (Carlini et al., 2023b; Nasr et al., 2023). The property of LFMs being universally attacked (Zou et al., 2023; Duan et al., 2023) and the jailbreak of LLMs (Huang et al., 2023; Chao et al., 2023; Wyllie et al., 2024) may also relate to the pre-training biases, but unidentified due to the lack of proper evaluation. These weaknesses of LFMs can usually not be found until they occur in practice, highlighting the necessity of more evaluation benchmarks from privacy and security.

Ethics and Bias. Biases in pre-training can also post vulnerabilities related to society science at the deployment of LFMs on downstream tasks, including misalignment (Wolf et al., 2023; Yang et al., 2023c), bias and fairness (Gallegos et al., 2023), and unethical content generation (Tokayev, 2023), affecting their reliability in critical applications like medical or financial systems. The misuse and unsafe deployment of LFMs (Mozes et al., 2023) also reflects the lack of comprehensive evaluation and proper metrics of them. Addressing the catastrophic inheritance of these biases is crucial for ensuring the reliability and fairness of LFMs.

2.4 Mitigation

Mitigating the impact of the biased pre-training data on LFMs, i,e, reducing g without full access and control of $f(\mathcal{D}_{up}, \mathcal{M}, \mathcal{A}_{up})$, is a complex and challenging task. The straightforward approach of identifying and filtering biases is difficult due to the need to maintain data diversity and quantity (Simig et al., 2022; Touvron et al., 2023a). Re-training LFMs can effectively mitigate specific biases, but requires significant computational resources and may introduce new issues (Longpre et al., 2023). Alternative strategies include unlearning techniques (Bourtoule et al., 2021; Xu et al., 2023a), allowing models to forget harmful biases (Bourtoule et al., 2021; Jang et al., 2022; Wu et al., 2024), and black-box methods that mitigate biases without full access to the model and data (Chen et al., 2024b; Oren et al., 2023). These approaches aim to balance bias mitigation with the practicalities and limitations of LFM tuning.

3 Challenges of Catastrophic Inheritance

We introduced and explored the concept and potential impacts of catastrophic inheritance as a significant challenge in the era of LFMs. In the following, we outline the primary difficulties on addressing it effectively and efficiently.

Availability Issue. The foremost obstacle is the assessment of pre-trained LFMs $\mathcal{M}$ and their pre-training data $\mathcal{D}_{\text{up}}$. While some models like LLaMA2 (Touvron et al., 2023b) and Mistral (Jiang et al., 2023) are open-source, their performance usually fall behind proprietary counterparts such as GPT-4 (OpenAI, 2023) and Gemini (Google, 2023). A recent notable effort, LLM-360 (Liu et al., 2023b), strives to provide more comprehensive open-source training details. The proprietary nature of models and datasets creates a "black box" environment for users and researchers, limiting our ability to identify biases and analyze the impacts. Additionally, the massive scale of these models demand substantial computational resources, even when they are open-source, making detailed exploration more challenging.

Evaluation Complexities. Another challenge is the evaluation of intelligence in LFMs (Chang et al., 2023). The evaluation should not only be conducted w.r.t. models in standard benchmarks but also in society with diverse human-AI interactions. Traditional benchmarks are inadequate for these assessments. The strong performance of LFMs is challenged due to potential data contamination (Roberts et al., 2023; Schaeffer, 2023; Jiang et al., 2024), inappropriate metrics (Schaeffer et al., 2023; Sun et al., 2023), or lack of standards (Zhu et al., 2023b; Lei et al., 2023). Furthermore, latent biases in LFMs mean that many potential harms remain invisible until they manifest in real-world outcomes, making proactive evaluation and mitigation more difficult.

Lack of understanding on LFMs. Third, while there is great advance in understanding the generalization of modern neural networks, the specific influence of pre-training data biases on this aspect, i.e., the format of g, particularly in real-world scenarios, is less explored. Most of the case studies in Table 1 are limited to well-structured and small-scale datasets, which cannot thoroughly represent the complex real-world data LFMs may encounter. Thus, applying existing theories to LFMs, interpreting their real-world behavior, and assessing their societal impact are still formidable tasks.

Trade-off in mitigation. Finally, despite some early attempts (Chen et al., 2024b; Lu et al., 2023), addressing pre-training biases involves a delicate balance between efficiency and effectiveness. Re-training LFMs on bias-free data is ideal, but not always feasible. Black-box methods can mitigate specific biases but might overlook or intensify others in different contexts. This creates difficulties for researchers in optimizing and mitigating catastrophic inheritance in downstream tasks without sacrificing performance or inadvertently increasing biases in other aspects.

4 Our UIM Framework

In this section, we present the UIM framework, as depicted in Figure 2, to address catastrophic inheritance. UIM calls for future research from three perspectives: understanding catastrophic inheritance from pre-training and evaluation in downstream tasks to figure out the trend of function g and f, interpreting the potential impacts and implications of biased pre-training data on downstream tasks both empirically and theoretically to find the form of g and f, and mitigating the adverse effects of biased pre-training data on downstream tasks to reduce g with the identified function relationship. It also serves as a general framework for studying CI, which we will show several existing works have already utlized it.

4.1 Understanding Catastrophic Inheritance

Fully understanding the impacts of catastrophic inheritance corresponds to finding the changes in both f and g from pre-training and downstream tasks respectively, including conducting empirical experiments and building novel evaluation metrics and benchmarks on various downstream tasks.

Probing into Effects at Pre-training and Downstream. The initial focus would be identifying the exact effects and the changes of them w.r.t. pre-training data biases at both the pre-training stage and downstream transferring stage. It is critical to study various types of pre-training data biases in this stage and find out the effects of such biases through large-scale experiments. As discussed earlier, the pre-training data biases not only shape the learning dynamics but consequently imprint on the model's behavior on downstream tasks (Nakkiran et al., 2019; Dar et al., 2021; Caballero et al., 2022). A particular aspect of interest is the relationship between these biases and scaling laws. We propose future research encompassing a comprehensive empirical investigation of different LFMs including CLIP (Radford et al., 2021), language models (Touvron et al., 2023a,b), and etc., under controlled and varying pre-training bias conditions. This investigation involves introducing different types and scales of synthetic and realistic biases into clean and controllable large-scale pre-training data, and illuminates the trend and changes in training dynamics, model behaviors, and generalization on downstream tasks. Changes of many other properties in f and g w.r.t biases is also worth studying, such as the expressive capacity (Zhang et al., 2016), transition from memorization to generalization (Power et al., 2022; Kumar et al., 2023; Davies et al., 2023), and structures of LFMs affected by the biases. Studying on the biases within controlled subset of concepts is also necessary (Feldman, 2020).

On the downstream side, we need to consider broader contexts to figure out the trend of g to biases. This includes evaluating the LFMs on diverse downstream datasets $\mathcal{D}_{\text{down}}$ of various domain and settings, such as tasks with imbalanced (Huang et al., 2016), noisy (Natarajan et al., 2013), few-shot (Wang et al., 2020), unlabeled (Wang et al., 2022), and ood (Hendrycks et al., 2021; Zhang et al., 2019) data, and different tuning algorithms $\mathcal{A}_{\text{down}}$, such as prompt strategies (Zhu et al., 2023c,d), linear probing, parameter-efficient tuning (He et al., 2021), and even full tuning. A more comprehensive evaluation not only facilitates identifying the trend of f and g but also the compositional relationship of them.

Evaluation Metrics/Benchmarks of LFMs. LFMs are difficult to evaluate holistically, not only because of their complex capability but also the lack of proper evaluation metrics and benchmarks. We advocate for the development of new evaluation metrics that go beyond traditional performance measures on downstream tasks. The metrics should incorporate different aspects of LFMs, such as (adversarial) robustness (Wang et al., 2023a), fairness (Du et al., 2020; Nanda et al., 2021), bias (Wu and Aji, 2023), security, and privacy (Yao et al., 2023). Furthermore, evaluations must address the potential misalignment between LFM behaviors and ethical or societal norms, ensuring that these powerful models act in ways that are beneficial and non-harmful to society at scale. It is important to design metrics that measure the explicit influence and memorization of biased samples (Feldman and Zhang, 2020; Carlini et al., 2022). Establishing novel and robust evaluation benchmarks is also vital, considering the prevalence of data contamination that obscures the true capabilities of LFMs. Dynamic evaluation protocols (Zhu et al., 2023b; Fan et al., 2023b) represent a

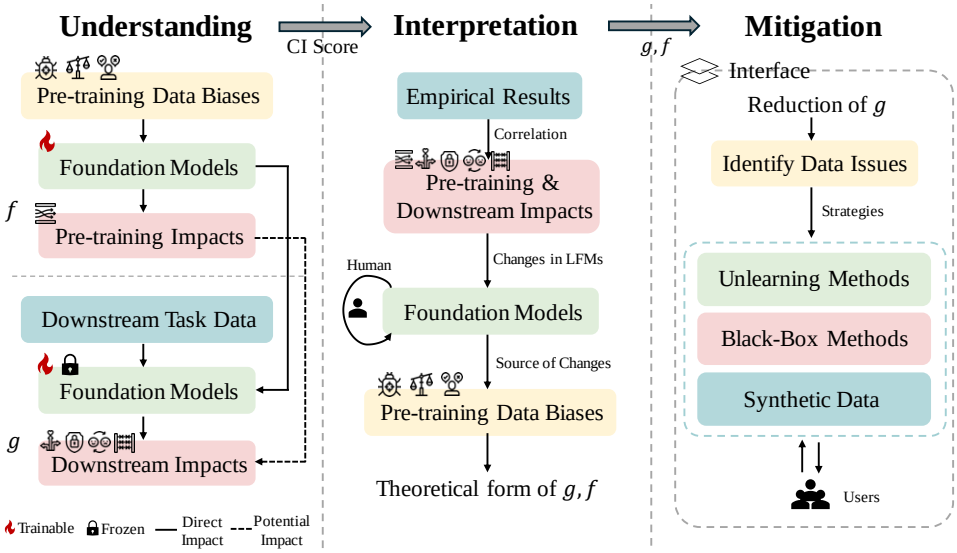

Figure 2: The UIM framework addressing catastrophic inheritance from understanding, interpretation, and mitigation.

promising direction. Future benchmarks should strive to generate rephrased, non-overlapping samples to counteract data contamination (Yang et al., 2023a). Additionally, collecting and annotating failure cases specifically arising from pre-training data biases at the deployment of LFMs will provide valuable insights for refining them. These dimensions are critical for understanding of catastrophic inheritance.

Understanding the Societal Impact. The assessment of catastrophic inheritance should be considered in an interdisciplinary fashion. It should broadly includes the human interaction with LLMs from the psychology aspects (Li et al., 2023b,a), and agents interaction within LLMs for studying society behaviors (Zhao et al., 2023b; Leng and Yuan, 2023; Park et al., 2023). We should also evaluate the effect of biases on critical applications of LFMs, such as medical and science (Singhal et al., 2022; Nejjar et al., 2023; Thirunavukarasu et al., 2023; Anderljung et al., 2023).

Several existing works have studied the understanding of LFMs from the perspective of pre-training data biases. For example, Zhu et al. (2023e) studied the data credibility issues in the pre-training of language models. Chen et al. (2024c,a) researched on the effects of pre-training corruption on CLIP and Diffusion models, respectively. Hu et al. (2024) studied the inheritance of adversarial examples from pre-training to downstream tasks of LFMs. Through a large-scale empirical study, Chen et al. (2024d) revealed the amplification effects of pre-training biases in diffusion models. All relevant works demonstrate that performing large-scale and well-controlled experiments about pre-training biases are essential to understanding their effects.

4.2 Interpreting the Impacts to Downstream Tasks

Interpreting why and how the malicious effects of pre-training data biases function in the downstream is crucial to address catastrophic inheritance, i.e., the exact form of g.

Empirical Interpretation of Malicious Effect. To empirically interpret the malicious effects of pre-training data biases, we need to conduct in-depth case studies and analyses on specific downstream applications. This involves examining the feature space using tools such as SVD, PCA, and T-SNE, both qualitatively and quantitatively. The singular values and vectors of the pre-trained feautres are often related to the transferability of generalization (Chen et al., 2024b; Xue et al., 2022; Chen et al., 2019). Jacobian matrix analysis is another perspective to explore transferability (Oymak et al., 2019), although its calculation in LFMs may require approximations (Yao et al., 2020). Such an empirical analysis also needs to be performed on a wide range of downstream tasks from different domains to assess g.

Theoretical Interpretation of Catastrophic Inheritance. In developing a theoretical interpretation of catastrophic inheritance in LFMs, we focus on frameworks that can precisely predict and articulate the observed bias inheritance from pre-training data to downstream tasks. This requires an in-depth examination of LFMs' internal mechanisms, particularly how they process and retain information from pre-training phases. Key to this exploration are concepts such as the balance between memorization and generalization (Zhang et al., 2016; Kumar et al., 2023; Zhu et al., 2024), the delineation of the memorization and generalization bounds (Kawaguchi et al., 2017), and also the theoretical evolution of LFM architectures during training. We aim to identify specific thresholds or critical points where pre-training biases critically influence these balances, similar in Nakkiran et al. (2019); Zhu et al. (2024). The theoretical frameworks help us to model the exact form of g.

Interpretation based on Human-AI Collaboration. The direction of adopting LFMs to correct themselves based on minimal and necessary human collaboration is also promising on interpreting the effects of pre-training biases (Jang, 2023). It involves design self-criticize (Wang et al., 2023b) and self-feedback loop based on human feedback to produce explanation and diagnosis (Gou et al., 2023) of LFMs themselves on the failures inherited from pre-training biases.

4.3 Mitigating Catastrophic Inheritance

Understanding and interpreting the malicious impacts on downstream tasks will help us design mitigation strategies.

Black-Box Tuning Methods. Black-box tuning is one of the most interesting methods for mitigating the malicious effects of the pre-training data biases on downstream tasks. These involve designing lightweight modules, such as additional layers, which can be applied to LFMs without altering their pre-trained weights. This approach is particularly intriguing due to its potential to remodel the feature space based on the biases identified in our empirical and theoretical analyses (Chen et al., 2019, 2024b). Similar methods have also been adopted in mitigating the adversarial noise of LFMs, especially in medical domain (Han et al., 2024). While parameter-efficient tuning methods share similarities with black-box approaches (Oh et al., 2023; Guo et al., 2023; Lin et al., 2023; Yu et al., 2023), they often require access to the internal structures or weights (He et al., 2021). Recently, Tong et al. (2024) have also tried tuning methods combining multiple models to mitigate the inheritance of a single model. Kim et al. (2024) proposed re-weighting methods along diffusion steps, specific to diffusion models, to learn unbiased diffusion models from biased datasets. Nonetheless, these black-box methods also present limitations, primarily due to the limited

scope of transformation they offer and the need to keep the pre-trained part of the model frozen. Future research will devise specialized regularization terms tailored to counteract the malicious effects of pre-training data biases, enhancing the effectiveness of these tuning methods.

Unlearning Methods. Machine unlearning techniques (Bourtoule et al., 2021; Xu et al., 2023a; Chen and Yang, 2023; Zhang et al., 2024a) can also be utilized to mitigate data biases prior to training in LFMs. The goal is to revise the LFM's knowledge to effectively forget, edit, and minimize the impact of biased data. Unlearning has also been adapted for diffusion models as demonstrated in (Wu et al., 2024), targeting the removal of learned biases. However, unlearned diffusion models have been found still generate unsafe contents recently (Zhang et al., 2023b). Chen and Yang (2023) introduced an efficient approach by integrating unlearning layers within transformer blocks to unlearn concepts in LFMs. Future developments should focus on designing novel methods requiring minimal interaction with the core structure and pre-training data of LFMs (Xiao et al., 2023) that effectively minimizes and unlearns the knowledge of certain types of biases from pre-training in LFMs. Addressing the trade-off between the bias mitigation at downstream and the performance degradation, as shown in existing works, is also an important question for future research.

Synthetic Data Tuning Methods. Synthetic data can also be utilized facilitate black-box tuning or unlearning methods in the situations where targeted biased data is inaccessible. The use of unbiased synthetic samples for further tuning, as demonstrated by Zhao et al. (2023a), can alleviate the effects of biases inherited from pre-training. Large diffusion models can be employed to generate the unbiased samples, which can be used to fine-tune LFMs (in black-box manners or for unlearning). Zheng et al. (2024) utilized synthetic data to solve the noisy label learning problem. Similarly, Seo et al. (2024) proposed to use synthetic c data for continual learning of LFMs. One promising research direction is to study to what extent the unbiased knowledge would be eliminated and how much unbiased data will be needed with synthetic data.

Pre-traning Data Curation and Pruning. Refining the pre-training dataset is a direct approach to mitigating biases, yet usually resource-intensive. Advanced tools for data inspection and documentation will be crucial in this process. Researchers will need to develop sophisticated metrics to effectively measure and balance the diversity, quality, and quantity of pre-training data, ensuring that the curated dataset is representative and free from harmful biases.

Designing Lifelong Interfaces. Novel platforms supporting lifelong updates of LFMs should be built, integrating the functions of identifying, understanding, interpreting, and mitigating the pre-training biases. After the stage of pre-training, the failures and misaligned behaviors of LFMs should be easily edited via the interaction of human on the platform to continually update the LFMs without re-training.

5 Conclusion

In this paper, we have identified an important yet neglected topic of LFMs, termed Catastrophic Inheritance, and delved into the multifaceted challenge of it, highlighting the critical need for understanding, interpreting, and mitigating the pre-training data biases. Our proposed UIM framework provides a comprehensive approach to understanding and ad-

dressing these issues. Through innovative methods such as black-box tuning, machine unlearning, synthetic data tuning, and pre-training data curation, we aim to advance the field in developing more robust, unbiased, and responsible LFMs. The future of LFMs depends on our ability to effectively manage and overcome the inherent biases in pre-training data. We hope this position paper inspires more research that contributes not only to the theoretical understanding of LFMs, but also to practical solutions to enhance their reliability and applicability in real-world scenarios.

6 Broader Impact

The research on Catastrophic Inheritance in LFMs addresses crucial concerns about the biases and limitations inherited from large-scale pre-training datasets. This work has significant implications across various domains. By highlighting the potential for these models to perpetuate and amplify biases, the study underscores the need for more responsible AI development and deployment. The proposed UIM framework aims to foster collaboration between machine learning and social sciences to better understand, interpret, and mitigate these biases. This interdisciplinary approach is essential for developing LFMs that are not only technically robust but also ethically sound and socially beneficial. We hope this position paper can guide future efforts in dataset curation, model training, and evaluation, ultimately contributing to the creation of fairer and more reliable AI systems.

References

Amro Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *ArXiv*, abs/2303.09540, 2023. URL https://api.semanticscholar.org/CorpusID:257557221.

Amro Abbas, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari S Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. *arXiv preprint arXiv:2401.04578*, 2024.

Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. *ArXiv*, abs/2110.02095, 2021. URL https://api.semanticscholar.org/CorpusID:238354065.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. (ab) using images and sounds for indirect instruction injection in multi-modal llms. *arXiv preprint arXiv:2307.10490*, 2023.

Mikhail Belkin, Daniel J. Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, 2:1167–1180, 2019. URL https://api.semanticscholar.org/CorpusID:81977297.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. Into the laions den: Investigating hate in multimodal datasets. *arXiv preprint arXiv:2311.03449*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Boston.com. An mit student asked ai to make her headshot more 'professional.' it gave her lighter skin and blue eyes. https://incidentdatabase.ai/cite/593/#r3264, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, H. Anderson, A. Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *ArXiv*, abs/2302.10149, 2023a. URL `https://api.semanticscholar.org/CorpusID:257038404`.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023b.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Hao Chen, Yujin Han, Diganta Misra, Xiang Li, Kai Hu, Difan Zou, Masashi Sugiyama, Jindong Wang, and Bhiksha Raj. Slight corruption in pre-training data makes better diffusion models. *arXiv preprint arXiv:2405.20494*, 2024a.

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Understanding and mitigating the label noise in pre-training on downstream tasks. In *International Conference on Learning Representations (ICLR)*, 2024b.

Hao Chen, Jindong Wang, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Learning with noisy foundation models. *arXiv preprint arXiv:2403.06869*, 2024c.

Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*, 2023.

Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, and Yuta Nakashima. Would deep generative models amplify bias in future models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10843, 2024d.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

Together Computer. Redpajama: an open dataset for training large language models, 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Corinna Coupette, Jyotsna Singh, and Holger Spamann. Simplify your law: Using information theory to deduplicate legal documents. *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 631–638, 2021. URL `https://api.semanticscholar.org/CorpusID:238259040`.

Y Dar, L Luzi, and RG Baraniuk. Frozen overparameterization: A double descent perspective on transfer learning of deep neural networks. *Procedia Computer Science*, 193:173–182, 2021.

Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*, 2023.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*, 2023.

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V Naidu, and Colin White. Forecastpfn: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 36, 2024.

Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023.

Rahim Entezari, Mitchell Wortsman, Olga Saukh, M Moein Shariatnia, Hanie Sedghi, and Ludwig Schmidt. The role of pre-training data in transfer learning. *arXiv preprint arXiv:2302.13602*, 2023.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*, 2023a.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, Yongfeng Zhang, and Libby Hemphill. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*, 2023b.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022.

Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? *arXiv preprint arXiv:2301.04644*, 2023.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Forbes. An mit student asked ai to make her headshot more 'professional.' it gave her lighter skin and blue eyes. https://www.forbes.com/sites/alexandralevine/2023/12/20/stable-diffusion-child-sexual-abuse-material-stanford-internet-observatory/?sh=4b4db3195f21, 2023.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*, 2024.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020a.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020b.

Google. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *ArXiv*, abs/2310.02664, 2023. URL `https://api.semanticscholar.org/CorpusID:263620137`.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Zixian Guo, Yuxiang Wei, Ming Liu, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Black-box tuning of vision-language models with effective gradient approximation. *arXiv preprint arXiv:2312.15901*, 2023.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

Xu Han, Linghao Jin, Xuezhe Ma, and Xiaofeng Liu. Light-weight fine-tuning method for defending adversarial noise in pre-trained medical vision-language models. *arXiv preprint arXiv:2407.02716*, 2024.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Anjun Hu, Jindong Gu, Francesco Pinto, Konstantinos Kamnitsas, and Philip Torr. As firm as their foundations: Can open-sourced foundation models be used to create adversarial examples for downstream tasks? *arXiv preprint arXiv:2403.12693*, 2024.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E Gonzalez, Koushik Sen, and Ion Stoica. Llm-assisted code cleaning for training accurate code generators. *arXiv preprint arXiv:2311.14904*, 2023a.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktaschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *ArXiv*, abs/2311.12786, 2023b. URL `https://api.semanticscholar.org/CorpusID:265308865`.

Eric Jang. Can llms critique and iterate on their own outputs? *evjang.com*, Mar 2023. URL `https://evjang.com/2023/03/26/self-reflection.html`.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *The WebConf*, 2024.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. *ArXiv*, abs/2202.06539, 2022. URL `https://api.semanticscholar.org/CorpusID:246823128`.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8), 2017.

Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il-Chul Moon. Training unbiased diffusion models from biased dataset. In *The Twelfth International Conference on Learning Representations*, 2024.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24, 2023a.

Hadas Kotek, Rikker Dockum, and David Q. Sun. Gender bias in llms, 2023b. URL `https://arxiv.org/abs/2308.14921`.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. The ethics of interaction: Mitigating security threats in llms. *arXiv preprint arXiv:2401.12273*, 2024.

Tanishq Kumar, Blake Bordelon, Samuel J Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. *arXiv preprint arXiv:2310.06110*, 2023.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. S3eval: A synthetic, scalable, systematic evaluation suite for large language models. *arXiv preprint arXiv:2310.15147*, 2023.

Yan Leng and Yuan Yuan. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*, 2023.

Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. *arXiv preprint arXiv:2312.16337*, 2023.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: Unveiling emotions in generative ai. *arXiv preprint arXiv:2312.11111*, 2023a.

Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv e-prints*, pages arXiv–2307, 2023b.

Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. Open the pandora's box of llms: Jailbreaking llms through representation engineering. *arXiv preprint arXiv:2401.06824*, 2024.

Yucheng Li. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation. *arXiv preprint arXiv:2309.10677*, 2023a.

Yucheng Li. An open source data contamination report for llama series models. *arXiv preprint arXiv:2310.17589*, 2023b.

Zihao Lin, Yan Sun, Yifan Shi, Xueqian Wang, Lifu Huang, Li Shen, and Dacheng Tao. Efficient federated prompt tuning for black-box large pre-trained models. *arXiv preprint arXiv:2310.03123*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023b.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

Wang Lu, Hao Yu, Jindong Wang, Damien Teney, Haohan Wang, Yiqiang Chen, Qiang Yang, Xing Xie, and Xiangyang Ji. Zoopfl: Exploring black-box foundation models for personalized federated learning. *arXiv preprint arXiv:2310.05143*, 2023.

Alexandra Luccioni and Joseph Viviano. What's in the box? an analysis of undesirable content in the common crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, 2021.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. Deciphering stereotypes in pre-trained language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.697. URL https://aclanthology.org/2023.emnlp-main.697.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:254877603.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023.

Marc Marone and Benjamin Van Durme. Data portraits: Recording foundation model training data. *arXiv preprint arXiv:2303.03919*, 2023.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D Griffin. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*, 2023.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021, 2019. URL https://api.semanticscholar.org/CorpusID:207808916.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. Llms for science: Usage for code generation and data analysis. *arXiv preprint arXiv:2311.16733*, 2023.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022.

Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24224–24235, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195, 2023.

OpenAI. Gpt-4 technical report, 2023.

Manfred Opper. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pages 922–925, 1995.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*, 2023.

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. *arXiv preprint arXiv:2401.12425*, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Alexandra Sasha Luccioni, Yacine Jernite, and Anna Rogers. The roots search tool: Data transparency for llms. *arXiv preprint arXiv:2302.14035*, 2023a.

Aleksandra Piktus, Odunayo Ogundepo, Christopher Akiki, Akintunde Oladipo, Xinyu Zhang, Hailey Schoelkopf, Stella Biderman, Martin Potthast, and Jimmy Lin. Gaia search: Hugging face and pyserini interoperability for nlp training data exploration. *arXiv preprint arXiv:2306.01481*, 2023b.

Alethea Power, Yuri Burda, Harrison Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *ArXiv*, abs/2201.02177, 2022. URL `https://api.semanticscholar.org/CorpusID:245769834`.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551, 2020.

Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ludwig Schmidt, and Ali Farhadi. On the connection between pre-training data diversity and fine-tuning robustness. *arXiv preprint arXiv:2307.12532*, 2023.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. *arXiv preprint arXiv:2312.03002*, 2023.

William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

Shahbaz Rezaei and Xin Liu. A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning. In *International Conference on Learning Representation (ICLR)*, 2020.

Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. Does progress on object recognition benchmarks improve real-world generalization? *arXiv preprint arXiv:2307.13136*, 2023.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. Data contamination through the lens of time. *arXiv preprint arXiv:2310.10628*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Rylan Schaeffer. Pretraining on the test set is all you need. *arXiv preprint arXiv:2309.08632*, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Minhyuk Seo, Diganta Misra, Seongwon Cho, Minjae Lee, and Jonghyun Choi. Just say the name: Online continual learning with category names only via data generation. *arXiv preprint arXiv:2403.10853*, 2024.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, and Yao Qin. Effective robustness against natural distribution shifts for models with different training data. *arXiv preprint arXiv:2302.01381*, 2023.

Ali Shirali and Moritz Hardt. What makes imagenet look unlike laion. *arXiv preprint arXiv:2306.15769*, 2023.

Daniel Simig, Tianlu Wang, Verna Dankers, Peter Henderson, Khuyagbaatar Batsuren, Dieuwke Hupkes, and Mona Diab. Text characterization toolkit (tct). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 72–87, 2022.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Khyathi Chandu, Jennifer Dumas, Li Lucy, Xinxi Lyu, et al. Dolma: An open corpus of 3 trillion tokens for language model pretraining research. *Allen Institute for AI, Tech. Rep*, 2023.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. *arXiv preprint arXiv:2311.17946*, 2023.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*, 2023.

David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical report, Stanford University, Palo Alto, CA, 2023, 2023. URL `https://doi.org/10.25740/kh752sm9123`.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S Morcos. D4: Improving llm pretraining via document de-duplication and diversification. *arXiv preprint arXiv:2308.12284*, 2023.

Kassym-Jomart Tokayev. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*, 8(9):17–33, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Vikrant Varma, Rohin Shah, Zachary Kenton, J'anos Kram'ar, and Ramana Kumar. Explaining grokking through circuit efficiency. *ArXiv*, abs/2309.02390, 2023a. URL `https://api.semanticscholar.org/CorpusID:261557247`.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. With great training comes great vulnerability: Practical attacks against transfer learning. In *27th USENIX security symposium (USENIX Security 18)*, pages 1281–1297, 2018.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023a.

Rui Wang, Hongru Wang, Fei Mi, Yi Chen, Ruifeng Xu, and Kam-Fai Wong. Self-critique prompting with large language models for inductive instructions. *arXiv preprint arXiv:2305.13733*, 2023b.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*, 2023c.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, Heli Qi, Zhen Wu, Yu-Feng Li, Satoshi Nakamura, Wei Ye, Marios Savvides, Bhiksha Raj, Takahiro Shinozaki, Bernt Schiele, Jindong Wang, Xing Xie, and Yue Zhang. Usb: A unified semi-supervised learning benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *ArXiv*, abs/2206.07682, 2022. URL `https://api.semanticscholar.org/CorpusID:249674500`.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023b.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: Training on synthetic data amplifies bias. *arXiv preprint arXiv:2403.07857*, 2024.

Guangxuan Xiao, Ji Lin, and Song Han. Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870*, 2023.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023a.

Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Cit: Curation in training for effective vision-language data. *arXiv preprint arXiv:2301.02241*, 2023b.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023c.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pages 24851–24871. PMLR, 2022.

Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9224, 2022.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023a.

Yu Yang, Aaditya K Singh, Mostafa Elhoushi, Anas Mahmoud, Kushal Tirumala, Fabian Gloeckle, Baptiste Rozière, Carole-Jean Wu, Ari S Morcos, and Newsha Ardalani. Decoding data quality via synthetic corruptions: Embedding-guided pruning of code data. *arXiv preprint arXiv:2312.02418*, 2023b.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023c.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *arXiv preprint arXiv:2312.02003*, 2023.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

Samuel Yu, Shihong Liu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2016. URL `https://api.semanticscholar.org/CorpusID:6212000`.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Jieyu Zhang, Bohan Wang, Zhengyu Hu, Pang Wei Koh, and Alexander Ratner. On the trade-off of intra-/inter-class diversity for supervised pre-training. *arXiv preprint arXiv:2305.12224*, 2023a.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024a.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*, 2024b.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*, 2023b.

Ziqi Zhang, Yuanchun Li, Jindong Wang, Bingyan Liu, Ding Li, Yao Guo, Xiangqun Chen, and Yunxin Liu. Remos: reducing defect inheritance in transfer learning via relevant model slicing. In *Proceedings of the 44th International Conference on Software Engineering*, pages 1856–1868, 2022.

Jiachen Zhao, Zhun Deng, David Madras, James Zou, and Mengye Ren. Learning and forgetting unsafe examples in large language models. *arXiv preprint arXiv:2312.12736*, 2023a.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv preprint arXiv:2310.17512*, 2023b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

Linghan Zheng, Hui Liu, Xiaojun Lin, Jiayuan Dong, Yue Sheng, Gang Shi, Zhiwei Liu, and Hongwei Chen. Top in chinese data processing: English code models. *arXiv preprint arXiv:2401.10286*.

Yuxiang Zheng, Zhongyi Han, Yilong Yin, Xin Gao, and Tongliang Liu. Can we treat noisy labels as accurate? *arXiv preprint arXiv:2405.12969*, 2024.

Tong Zhou, Yubo Chen, Pengfei Cao, Kang Liu, Jun Zhao, and Shengping Liu. Oasis: Data curation and assessment system for pretraining of large language models. *ArXiv*, abs/2311.12537, 2023. URL `https://api.semanticscholar.org/CorpusID:265308678`.

Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. *arXiv preprint arXiv:2310.08106*, 2023a.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*, 2023b.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023c.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. Promptbench: A unified library for evaluation of large language models. *arXiv preprint arXiv:2312.07910*, 2023d.

Xuekai Zhu, Yao Fu, Bowen Zhou, and Zhouhan Lin. Critical data size of language models from a grokking perspective. *arXiv preprint arXiv:2401.10463*, 2024.

Zhaowei Zhu, Jialu Wang, Hao Cheng, and Yang Liu. Unmasking and improving data credibility: A study with datasets for training harmless language models. *arXiv preprint arXiv:2311.11202*, 2023e.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.