FAIRSISA: ENSEMBLE POST-PROCESSING TO IMPROVE FAIRNESS OF UNLEARNING IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training large language models (LLMs) is a costly endeavour in terms of time and computational resources. The large amount of training data used during the unsupervised pre-training phase makes it difficult to verify all data and, unfortunately, undesirable data may be ingested during training. Re-training from scratch is impractical and has led to the creation of the *unlearning* discipline where models are modified to "unlearn" undesirable information without retraining. However, any modification can alter the behaviour of LLMs, especially on key dimensions such as *fairness*. This is the first work that examines this interplay between unlearning and fairness for LLMs. In particular, we focus on a popular unlearning framework known as SISA [Bourtoule et al., 2021], which creates an ensemble of models trained on disjoint shards. We evaluate the performance-fairness trade-off for SISA, and empirically demsontrate that SISA can indeed reduce fairness in LLMs. To remedy this, we propose post-processing bias mitigation techniques for ensemble models produced by SISA. Through experimental results, we demonstrate the efficacy of our post-processing framework called *FairSISA*.

1 INTRODUCTION

Modern machine learning models, especially large language models (LLMs), employ significantly large model architectures and train on massive datasets. At this scale, it is infeasible to properly curate the training data, and sensitive or undesirable information (e.g., personally identifiable information, copyrighted material or toxic text) may be ingested by the model during training Carlini et al. (2021); Lehman et al. (2021); Carlini et al. (2023a;b). Moreover, when models are trained on data collected from individual users, e.g., medical data, some users may request their data to be deleted following the right to be forgotten provided by recent privacy legislation Voigt & Von dem Bussche (2017); Pardau (2018); Act (2000).

Motivated by the above scenarios, *machine unlearning* has emerged as a subfield of machine learning. The goal of machine unlearning is to remove the influence of a specific subset of training examples from a trained model, while maintaining the performance of the model. A straightforward machine unlearning method is to retrain the model on an updated training set that excludes the samples to be removed. However, retraining deep models, especially LLMs, from scratch is infeasible due to the exorbitant computational costs. Several unlearning techniques have recently been proposed to tackle this challenge by efficiently removing the influence of the data to be unlearned (see Appendix A for an overview).

Even though machine unlearning has recently received significant research attention, implications of unlearning on other crucial aspects such as fairness have been scantly explored. Fairness is especially critical for language models, since these models are embedded in a variety of applications including call centers and other question-answer applications where the output my jeopardize people's chances to obtain services or may lead to unfair treatment. Several works have demonstrated the bias of language models, see, e.g., Bolukbasi et al. (2016); Borkan et al. (2019); Hutchinson et al. (2020); de Vassimon Manela et al. (2021); Baldini et al. (2022). However, little understanding on the effects of unlearning on the overall fairness of the model.

In this work, we evaluate the fairness of LLMs trained using a popular unlearning framework, called Sharded, Isolated, Sliced, and Aggregated (SISA) Training Bourtoule et al. (2021). SISA framework partitions the training data into disjoint shards, and trains a constituent model on each shard,

thus creating an ensemble of models. During inference, predictions from the constituent models are aggregated, typically using majority voting. During unlearning, only impacted constituent model needs to be retrained on much smaller shard, resulting in significant speed ups. Our choice of SISA among unlearning techniques is motivated by the following reasons. First, SISA is an *exact* unlearning framework with certifiable guarantees. In contrast, several unlearning methods provide only *approximate* unlearning (see Sec. 2 for details). Second, SISA framework can be applied to a wide variety of model architectures including transformer-based language models. On the contrary, several unlearning methods, especially with certified guarantees, are not applicable to language models (e.g., Sekhari et al. (2021); Warnecke et al. (2023)). Overall SISA can comply with regulations that require data removal at the requests of the data owner.

To improve the fairness of LLMs trained using SISA, we propose to apply *post-processing techniques* for bias mitigation. Unlike *pre-processing* and *in-processing* bias mitigation techniques, post-processing techniques do not require any modification of the training data or model training procedures, and they only modify the model output (see Bellamy et al. (2019) for a summary). Given the massive costs associated with training large LLMs and the vast amount of unlabeled data used during this process, we argue that post-processing approaches can minimize the environmental impact of re-training LLMs.

Contributions: To the best of our knowledge, this is the first work to explore the effects of unlearning on the fairness of LLMs. We outline our contributions below.

- We study the accuracy-fairness trade-off of LLMs trained using the SISA unlearning framework by focusing on the task of toxic text classification. We measure model bias in terms of *group fairness* using the notion of *equalized odds*, by following the setup in Baldini et al. (2022). We empirically demonstrate that the SISA framework can produce models that are less fair.
- We investigate post-processing bias mitigation techniques (adapted from Hardt et al. (2016)) in the context of SISA ensembles to improve the accuracy-fairness trade-off. We adapt the post-processing method from Hardt et al. (2016) to design three methods that can handle model ensembles. We prove that the third method generalizes the post-processing optimization problem in Hardt et al. (2016) for ensemble of models and produces an optimal fair predictor for ensemble of models, which is our key theoretical contribution and can be of independent interest outside of SISA unlearning. We empirically evaluate the three post-processing methods for SISA on two state-of-the-art LLM architectures for two datasets.

2 PRELIMINARIES

Machine Unlearning: Let $\mathcal{D}$ be a fixed training dataset consisting of N samples, i.e., $\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, where each $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}$. Let us abstract out the training process using a (randomized) algorithm $\mathcal{A}$ that trains on $\mathcal{D}$ and outputs a model $\mathbf{w} \in \mathcal{W}$, where $\mathcal{W} \subseteq \mathbb{R}^d$ denotes the parameter space of a hypothesis class. Note that randomness in $\mathcal{A}$ induces a probability distribution over the models in the parameter space.

An unlearning algorithm $\mathcal{U}$ takes as input the trained model $\mathbf{w} = \mathcal{A}(\mathcal{D})$, data to be unlearned $\mathcal{D}^u \subset \mathcal{D}$ and retain dataset $\mathcal{D}^r = \mathcal{D} \setminus \mathcal{D}^u$, and outputs a new model $\mathbf{w}^u$, i.e., $\mathbf{w}^u = \mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}^u, \mathcal{D}^r)$. *Exact unlearning* methods essentially ensure that the distribution of the unlearned model $\mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}^u, \mathcal{D}^r)$ is perfectly indistinguishable from that of the retrained model $\mathcal{A}(\mathcal{D}^r)$ Bourtoule et al. (2021). On the other hand, *approximate unlearning* methods ensure that the distributions of unlearned and retrained models are stochastically indistinguishable, where stochastic indistinguishability is typically characterized by using notions similar to differential privacy Guo et al. (2019); Sekhari et al. (2021); Warnecke et al. (2023).

Sharded, Isolated, Sliced, and Aggregated (SISA) Training: Bourtoule et al. (2021) proposed SISA, an exact unlearning method that reduces the computational overhead associated with retraining from scratch. The SISA framework randomly divides the training dataset $\mathcal{D}$ into S disjoint shards $\mathcal{D}_1, \ldots, \mathcal{D}_S$ of approximately equal size. During training, for each shard $\mathcal{D}_k$, a *constituent model*, denoted as M_k, is trained. The data in each shard is further partitioned into R disjoint slices, where each constituent model is trained incrementally on each slice and the model parameters after training on a slice are saved. At inference time, S individual predictions from the constituent models are aggregated, typically, through majority voting (similar to ensemble methods Dietterich (2000)).

When one or more data samples need to be unlearned, only the constituent models corresponding to the shards that contain the data sample(s) are retrained. More specifically, the slice containing the data sample(s) to be unlearned and the following slices in the same shard need to be retrained. In other words, the last saved checkpoint before including the slice containing the data sample(s) to be unlearned can be used as a starting point. This provides significant speed ups over conventional retraining of a single model.

Fairness for Toxic Text Classification: We consider the task of toxic text classification, and measure model bias in terms of *group fairness* Chouldechova & Roth (2018) by following the setup in Baldini et al. (2022). In particular, we consider certain topics, such as religion or race, as sensitive. If a text sample mentions one of the sensitive topics (e.g., religion), we say that it belongs to a *sensitive group*; otherwise, to the complementary group (no religion). We analyze the fairness of toxic text prediction in the presence or absence of sensitive information (e.g., religion or race), with the goal that the performance of a fair predictor should not be influenced by these sensitive topics.

While there are several notions of group fairness, e.g., demographic parity (see Verma & Rubin (2018); Czarnowska et al. (2021)), we consider the notion of *equalized odds* Hardt et al. (2016). Essentially, equalized odds requires that the model output conditioned on the true label to be independent of the sensitive attribute. More formally, let Y denote the true label (e.g., toxic text), X denote the features, and A denote the sensitive attribute (e.g., religion or race). Let $\hat{Y} = f_{\mathbf{w}}(X, A)$ be the model output, denoted as the *predictor*. Equalized odds requires that the model predictor $\hat{Y}$ has equal *true positive rates* and *false positive rates* across the privileged and unprivileged groups, satisfying the following constraint:

$$\Pr\left(\hat{Y} = 1 \mid A = 0, Y = y\right) = \Pr\left(\hat{Y} = 1 \mid A = 1, Y = y\right), \quad y \in \{0, 1\}. \tag{1}$$

Baseline Post-Processing Method for Fairness: To improve the model fairness without retraining, we explore the use of post-processing methods. We build on the post-processing method proposed in Hardt et al. (2016), who originally proposed the notion of equalized odds. We denote the method as *HPS*, using the last names of the authors.

The HPS method constructs a *derived predictor* $\tilde{Y}$, which only depends on the predicted label $\hat{Y}$ and the sensitive attribute A, and satisfies equalized odds while minimizing classification loss. Specifically, let $\ell : \{0, 1\}^2 \to \mathbb{R}$ denote a loss function that takes a pair of labels and returns a real number. Let us define $p_{ya} = \Pr\left(\tilde{Y} = 1 \mid \hat{Y} = y, A = a\right)$. Then, the HPS method constructs $\tilde{Y}$ by solving the following optimization problem:

$$\begin{aligned} \min_{p_{ya}} \quad & \mathbb{E}\left[\ell(\tilde{Y}, Y)\right] \\ \text{s.t.} \quad & \Pr\left(\tilde{Y} = 1 \mid A = 0, Y = 0\right) = \Pr\left(\tilde{Y} = 1 \mid A = 1, Y = 0\right), \\ & \Pr\left(\tilde{Y} = 1 \mid A = 0, Y = 1\right) = \Pr\left(\tilde{Y} = 1 \mid A = 1, Y = 1\right), \\ & 0 \le p_{ya} \le 1. \end{aligned} \tag{2}$$

One can show that the above optimization problem is a linear program in four variables $\{p_{ya} : y \in \{0, 1\}, a \in \{0, 1\}\}$ Hardt et al. (2016). We denote the derived predictor obtained by solving the above optimization problem as $\text{HPS}(\hat{Y})$. Next, we adapt the HPS method to design post-processing methods for the ensemble of models produced by SISA.

3 FairSISA: Ensemble Post-Processing for SISA

Let $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S$ denote the predictions from the SISA constituent models. We consider three ways to perform post-processing for SISA.

Aggregate then post-process: The most natural way to apply post-processing to SISA is after aggregating the predictions from the constituent models. We focus on the majority voting aggregation rule, since it is demonstrated to perform well Bourtoule et al. (2021). We denote majority voting as

$$\textsc{Maj}\left(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S\right) = \arg\max_{y \in \{0,1\}} n_y, \quad \text{where } n_y = \left|\{i \in [S] : \hat{Y}_i = y\}\right|. \tag{3}$$

Then, the derived predictor obtained by first aggregating and then post-processing can be defined as $\text{HPS}\left(\text{MAJ}\left(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S\right)\right)$.

Post-process then aggregate: Another natural way to apply post-processing to SISA is to first post-process the label from each constituent model and then aggregate the post-processed predictions. Again, focusing on the majority voting aggregation rule, the derived predictor obtained by first post-processing and then aggregating can be defined as $\text{MAJ}\left(\text{HPS}(\hat{Y}_1), \text{HPS}(\hat{Y}_2), \ldots, \text{HPS}(\hat{Y}_S)\right)$.

Ensemble post-processing: Instead of aggregating the predictions before or after post-processing with a specific aggregation rule (such as majority voting), we design a post-processing method that can inherently aggregate the predictions. In particular, we generalize the HPS optimization problem to handle ensemble predictions. Recall that $\ell : \{0,1\}^2 \to \mathbb{R}$ denotes a loss function that takes a pair of labels and returns a real number. For a length-S binary vector $\bar{y} \in \{0,1\}^S$ and $a \in \{0,1\}$, let us define $p_{\bar{y}a} = \Pr\left(\tilde{Y} = 1 \mid \hat{Y}_1 = \bar{y}_1, \hat{Y}_2 = \bar{y}_2, \ldots, \hat{Y}_S = \bar{y}_S, A = a\right)$. We propose an ensemble post-processing method that constructs $\tilde{Y}$ by solving the following optimization problem:

$$\begin{aligned} \min_{p_{\bar{y}s}} \quad & \mathbb{E}\left[\ell(\tilde{Y}, Y)\right] \\ \text{s.t.} \quad & \Pr\left(\tilde{Y} = 1 \mid A = 0, Y = 0\right) = \Pr\left(\tilde{Y} = 1 \mid A = 1, Y = 0\right), \\ & \Pr\left(\tilde{Y} = 1 \mid A = 0, Y = 1\right) = \Pr\left(\tilde{Y} = 1 \mid A = 1, Y = 1\right), \\ & 0 \le p_{\bar{y}s} \le 1. \end{aligned} \tag{4}$$

Next, we show that the above optimization problem is a linear program, which produces an optimal derived predictor for the ensemble of models. The proof is deferred to Appendix B

Proposition 1 *The optimization problem in equation 4 is a linear program in 2^{S+1} variables $p_{\bar{y}a}$, whose coefficients can be computed from the joint distribution of $(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S, A, Y)$. Further, its solution is an optimal equalized odds predictor derived from $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S$ and A.*

4 EVALUATION

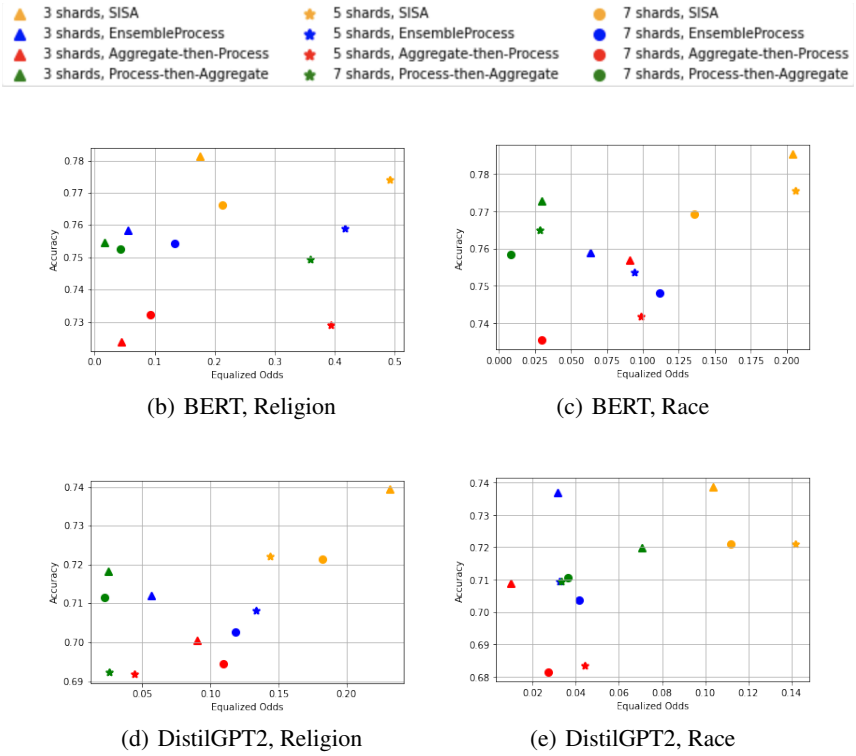
We perform empirical evaluation on using three state-of-the-art models (BERT, DistilGPT2, GPT2) on a representative dataset (HateXplain). For the sake of space, in the main paper, we discuss empirical results for BERT and DistilGPT2. Additional details and empirical evaluations are in Appendix C. HateXplain is a benchmark hate speech dataset which consists of 20K posts from Twitter and Gab Mathew et al. (2021). The dataset has fine-grained annotations for religion, race, and gender. We use coarse-grained groups as sensitive groups (e.g., mention of any religion) as opposed to the finer-grained annotations (e.g., Hindu), similar to Baldini et al. (2022). This is because, for the datasets we used, most subgroups account for significantly less proportion of the data, and there is considerable overlap between subgroups. We focus on two sensitive attributes: religion and race. We combine the annotations for offensive and hate speech into one class of toxic text, similar to Baldini et al. (2022).

First, we investigate how SISA training procedure influences the performance-fairness relationship by considering $S = 1, 3, 5$, and 7 shards[1]. Note that $S = 1$ shard corresponds to the conventional single model fine-tuning paradigm. In Figure 1, we demonstrate the performance as measured by accuracy on y-axis (higher accuracy is better) and the group fairness as measured by equalized odds (EO) on the x-axis (lower EO is better). We observe that, for both the models and sensitive attributes, the accuracy generally decreases with the number of shards, which is consistent with the observation in Bourtoule et al. (2021) for image-domain data. In contrast, EO values vary widely for different number of shards, and the SISA framework can indeed degrade the fairness (with higher EO values) for both the models and sensitive attributes. These results strongly suggest that it is important to investigate bias mitigation methods for the SISA framework.

[1]We do not consider the *slicing* component of SISA, because, unlike smaller models studied in Bourtoule et al. (2021), LLMs incur prohibitively large storage cost for saving model checkpoints for each slice.

(a) Religion (b) Legend (c) Race

Figure 1: Accuracy-fairness trade-off for SISA framework.



(b) BERT, Religion (c) BERT, Race

(d) DistilGPT2, Religion (e) DistilGPT2, Race

Figure 2: Comparison of post-processing methods for SISA.

Next, we compare three post-processing methods for bias mitigation from Section 3 for the SISA framework. In Figure 2, for each model and sensitive attribute, we plot accuracy vs. equalized odds (EO). Amongst the three methods, *Post-process then Aggregate* method achieves the best trade-off between the accuracy and EO, whereas *Aggregate then Post-Process* method achieves the worst trade-off between the accuracy and EO. The *Ensemble Post-Process* method generally achieves the highest accuracy for a moderate EO, which is consistent with the theory that the method is optimal in terms of accuracy (the objective function of the optimization problem equation 4).

Conclusion: We investigated an interplay between unlearning and fairness for LLMs by focusing on a popular unlearning framework called SISA Bourtoule et al. (2021). We empirically demonstrated that SISA can indeed reduce fairness in LLMs. As a solution, we proposed three post-processing bias mitigation techniques for ensemble models produced by SISA. We theoretically showed that one of the methods generalizes the optimization problem from Hardt et al. (2016) for ensemble models and produces an optimal derived predictor. We empirically demonstrated the efficacy of our post-processing techniques for SISA.

REFERENCES

Privacy Act. Personal information protection and electronic documents act. *Department of Justice, Canada. Full text available at http://laws. justice. gc. ca/en/P-8.6/text. html*, 2000.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2245–2262, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pp. 642–653, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pp. 491–500, New York, NY, USA, 2019. Association for Computing Machinery.

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)*, pp. 141–159. IEEE, 2021.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, August 2021. ISBN 978-1-939133-24-3.

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*, 2023a.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical, 2023b.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018. URL `http://arxiv.org/abs/1810.08810`.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 11 2021.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2232–2242, Online, April 2021. Association for Computational Linguistics.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision*, pp. 383–398. Springer, 2020b.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. *arXiv preprint arXiv:2010.10981*, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501, Online, July 2020. Association for Computational Linguistics.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Korbinian Koch and Marcus Soll. No matter how you slice it: Machine unlearning with sisa comes at the expense of minority classes. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 622–637. IEEE, 2023.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pp. 853–862, 2018.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 946–959. Association for Computational Linguistics, June 2021.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Alex Oesterling, Jiaqi Ma, Flavio P Calmon, and Hima Lakkaraju. Fair machine unlearning: Data removal while mitigating disparities. *arXiv preprint arXiv:2307.14754*, 2023.

Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. *arXiv preprint arXiv:2109.13398*, 2021.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pp. 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society, 2023.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio Calmon. Optimized score transformation for fair classification. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1673–1683. PMLR, 26–28 Aug 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

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

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Dawen Zhang, Shidong Pan, Thong Hoang, Zhenchang Xing, Mark Staples, Xiwei Xu, Lina Yao, Qinghua Lu, and Liming Zhu. To be forgotten or to be fair: Unveiling fairness implications of machine unlearning methods. *arXiv preprint arXiv:2302.03350*, 2023.

Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1):1–36, 2020.

APPENDIX

A RELATED WORK

Machine Unlearning. Cao & Yang (2015) were the first to introduce the notion of machine unlearning. Machine unlearning approaches can be divided in two broad categories (i) exact machine unlearning (e.g., retraining from scratch, SISA Bourtoule et al. (2021)) and (ii) approximate machine unlearning Graves et al. (2020); Izzo et al. (2021); Ginart et al. (2019); Golatkar et al. (2020a;b); Thudi et al. (2021). For details, see recent surveys Nguyen et al. (2022); Xu et al. (2023).

AI Fairness. The goal of AI fairness is to identify and eliminate algorithmic bias from machine learning models. This bias can arise from the difference between individuals or groups with respect to a sensitive attribute (e.g., race, gender, status, etc.). Fairness in machine learning has been widely studied Bellamy et al. (2019); Zhang et al. (2020); Biswas & Rajan (2020); Dwork et al. (2012); Mehrabi et al. (2021). Several studies have proposed bias mitigation techniques, which can be grouped in (i) pre-processing Calmon et al. (2017); Krasanakis et al. (2018), (ii) in-processing Corbett-Davies et al. (2017); Zhang et al. (2018), and (iii) post-processing techniques Hardt et al. (2016); Pleiss et al. (2017); Wei et al. (2020). In this work, we focus on post-processing methods because they are directly applied to outputs of the trained model, without requiring to modify the model.

Fairness vs Machine Unlearning. Even though AI fairness has been widely studied and various machine unlearning techniques have been developed, the literature still lacks works studyng the impact of machine unlearning techniques. Zhang et al. (2023) study fairness properties of different unlearning methods applied to small models. Their results on various tabular datasets show that unlearning can increase disparity. Oesterling et al. (2023) present the first provably fair unlearning method. They conduct extensive experiments on tabular datasets showing its efficacy. However, this method is restricted to convex functions. Koch & Soll (2023) analyze the performance of SISA models in imbalanced datasets. To the best of our knowledge, no work has investigated the impact of SISA models on fairness for LLMs and studied the application of post-processing methods to mitigate it.

B PROOF OF PROPOSITION 1

By the definition of derived predictor $\tilde{Y}$ Hardt et al. (2016), it can only depend on ensemble output $\hat{Y}_1, \ldots, \hat{Y}_S$ and A. Since these variables are binary, the predictor $\tilde{Y}$ is completely defined by the 2^{S+1} parameters in $[0, 1]$ corresponding to the probabilities $p_{\bar{y}a} = \Pr\left(\tilde{Y} = 1 \mid \hat{Y}_1 = \bar{y}_1, \hat{Y}_2 = \bar{y}_2, \ldots, \hat{Y}_S = \bar{y}_S, A = a\right)$.

Next, we show that the objective function is a linear function in these parameters. For simplicity of notation, we define, $p^y_{\bar{y}a} = \Pr\left(\tilde{Y} = y \mid \hat{Y}_1 = \bar{y}_1, \hat{Y}_2 = \bar{y}_2, \ldots, \hat{Y}_S = \bar{y}_S, A = a\right)$. Also, let us denote $\hat{Y}_e$ as the vector of predictions from the S constituent models, i.e., $\hat{Y}_e = [\hat{Y}_1\ \hat{Y}_2\ \ldots\ \hat{Y}_S]$.

$$
\begin{aligned}
\mathbb{E}[\ell(\tilde{Y}, Y)] &= \sum_{y', y \in \{0,1\}, y' \neq y} \Pr\left(\tilde{Y} = y', Y = y\right) \\
&= \sum_{y', y \in \{0,1\}, y' \neq y} \sum_{a \in \{0,1\}, y'' \in \{0,1\}^S} \Pr\left(\tilde{Y} = y', Y = y \mid \hat{Y}_e = y'', A = a\right) \Pr\left(\hat{Y}_e = y'', A = a\right) \\
&= \sum_{y', y, y' \neq y} \sum_{a, y''} \Pr\left(\tilde{Y} = y' \mid \hat{Y}_e = y'', A = a\right) \\
&\qquad \Pr\left(Y = y \mid \hat{Y}_e = y'', A = a\right) \Pr\left(\hat{Y}_e = y'', A = a\right) \\
&= \sum_{y', y, y' \neq y} \sum_{a, y''} p^y_{\bar{y}a} \Pr\left(Y = y, \hat{Y} = y'' \mid A = a\right) \Pr\left(A = a\right) \qquad (5)
\end{aligned}
$$

Since all probability in the last line above that do not involve $\tilde{Y}$ can be computed from the joint distribution, it follows that the objective function is linear in $p_{\bar{y}a}$.

To show that the equalized odds constraints are linear in $p_{\bar{y}a}$, consider the true positive rate (TPR) constraint in the equalized odds definition:

$$
\begin{aligned}
&\Pr\left(\tilde{Y}=1 \mid Y=1, A=a\right) \\
&= \sum_{\bar{y}\in\{0,1\}^S} \Pr\left(\tilde{Y}=1, \hat{Y}_e=\bar{y} \mid Y=1, A=a\right) \\
&= \sum_{\bar{y}} \Pr\left(\hat{Y}_e=\bar{y} \mid Y=1, A=a\right) \Pr\left(\tilde{Y}=1 \mid Y=1, \hat{Y}_e=\bar{y}, A=a\right) \\
&= \sum_{\bar{y}} \frac{\Pr\left(\hat{Y}_e=\bar{y}, Y=1 \mid A=a\right)}{\Pr\left(Y=1 \mid A=a\right)} \Pr\left(\tilde{Y}=1 \mid Y=1, \hat{Y}_e=\bar{y}, A=a\right) \\
&= \sum_{\bar{y}} \frac{\Pr\left(\hat{Y}_e=\bar{y}, Y=1 \mid A=a\right)}{\Pr\left(Y=1 \mid A=a\right)} \Pr\left(\tilde{Y}=1 \mid \hat{Y}_e=\bar{y}, A=a\right) \\
&= \sum_{\bar{y}} \frac{\Pr\left(\hat{Y}_e=\bar{y}, Y=1 \mid A=a\right)}{\Pr\left(Y=1 \mid A=a\right)} p_{\bar{y}a} \qquad (6)
\end{aligned}
$$

All probabilities in the last line above that do not involve $\tilde{Y}$ can be computed from the joint distribution, and it follows that the TPR is linear in $p_{\bar{y}a}$. Similarly, we can show that the false positive rate portion of the EO constraints is also linear in $p_{\bar{y}a}$.

Therefore, both the objective function and the constraints of equation 4 are linear in $p_{\bar{y}a}$, which completes the proof of the first part of Proposition 1. The optimality $\tilde{Y}$ follows from the fact that the optimal solution of equation 4 minimizes the loss while satisfying equalized odds, which completes the proof of the second part of Proposition 1.

C Additional Empirical Evaluations

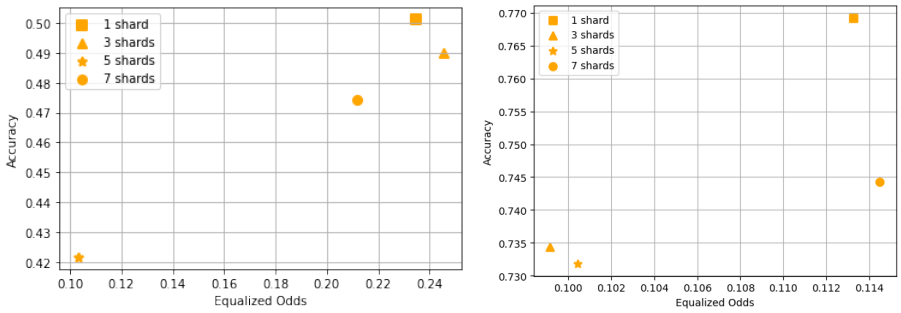
C.1 Details on Evaluation Parameters

We use the Hugging Face implementation of Transformers Wolf et al. (2020) and the corresponding implementations for language models. We use the text sequence classifier without any modifications to increase reproducibility. We fine-tune models for $[S, S+1, S+2]$ epochs, where S is the number of shards ($S = 1$ for the baseline single model case), and choose the parameter based on validation accuracy. To tackle any variance, we run experiments for two random seeds and report the average results.

C.2 Evaluation of GPT2

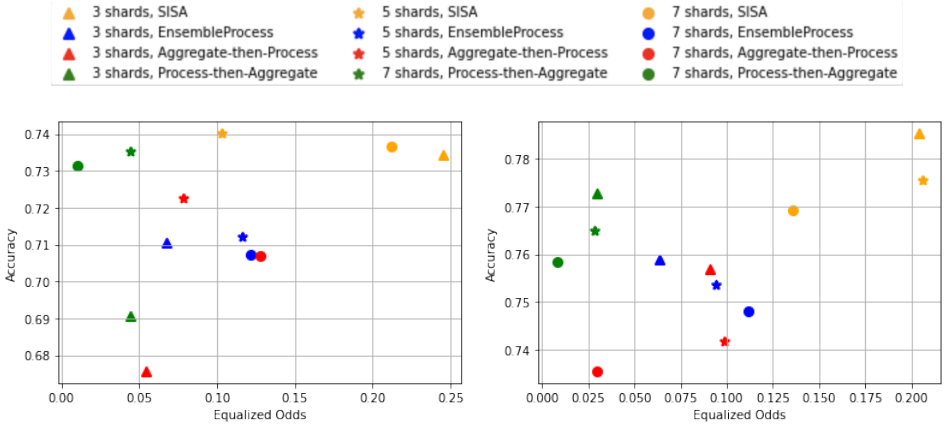
In Figure 3, we demonstrate the performance as measured by accuracy on y-axis (higher accuracy is better) and the group fairness as measured by equalized odds (EO) on the x-axis (lower EO is better). We observe similar trends as for BERT and DistilGPT2. For both sensitive attributes, the accuracy generally decreases with the number of shards, whereas, EO values vary for different number of shards. For the religion attribute, the SISA framework can degrade the fairness (with higher EO values) for GPT2.

Next, we compare three post-processing methods for bias mitigation from Section 3 for the SISA framework. In Figure 4, we plot accuracy vs. equalized odds (EO). Again, we observe similar trends as for BERT and DistilGPT2. Amongst the three methods, *Post-process then Aggregate* method achieves the best trade-off between the accuracy and EO.



(a) Religion (b) Race

Figure 3: Accuracy-fairness trade-off for SISA framework for GPT2.



(b) GPT2, Religion (c) GPT2, Race

Figure 4: Comparison of post-processing methods for SISA for GPT2.

C.3 One Fair Shard

Next, we study the scenario where one shard is fair and the others are unfair. To obtain a fair shard, we get an equal number of data samples from each possible combination of the values of the sensitive attribute (religion or non-religion) and the labels (toxic or normal). We split the remaining data samples randomly between the other shards. We note that, if the unlearning likelihood of individual data sample is known or can be estimated, SISA can place data samples with a high unlearning likelihood on designated shards. This can indeed result in some shards being unfair.

We compare the post-processing methods for $S = 3$ and 5 shards for the sensitive attribute of religion in Figure 5. First, we observe that the EO values without any post-processing are much larger that the case of random splitting. Post-processing methods significantly reduce model bias, and the *Ensemble Post-processing* methods achieves highest accuracy and substantially low EO, which is consistent with its theoretical optimality.

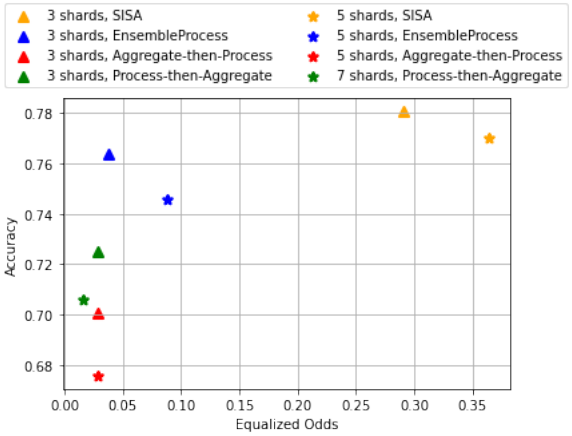


Figure 5: Comparison of post-processing methods for SISA for BERT when one shard is fair and the others are unfair.