Characterizing Stereotypical Bias from Privacy-preserving Pre-Training

Stefan Arnold and **Rene Gröbner** and **Annika Schreiner**
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
(stefan.st.arnold, rene.edgar.gröbner, annika.schreiner)@fau.de

Abstract

Differential Privacy (DP) can be applied to raw text by exploiting the spatial arrangement of words in an embedding space. We investigate the implications of such text privatization on Language Models (LMs) and their tendency towards stereotypical associations. Since previous studies documented that linguistic proficiency correlates with stereotypical bias, one could assume that techniques for text privatization, which are known to degrade language modeling capabilities, would cancel out undesirable biases. By testing BERT models trained on texts containing biased statements primed with varying degrees of privacy, our study reveals that while stereotypical bias generally diminishes when privacy is tightened, text privatization does not uniformly equate to diminishing bias across all social domains. This highlights the need for careful diagnosis of bias in LMs that undergo text privatization.

1 Introduction

Language Models (LMs) (Devlin et al., 2019; Radford et al., 2019) are trained on large corpora of text that may contain confidential information. Since such information can be recovered from word embeddings (Song and Raghunathan, 2020; Thomas et al., 2020) and language models (Carlini et al., 2019; Nasr et al., 2023), privacy emerged as an active concern for building trust and complying with stringent regulations on privacy protection.

To protect against unintended disclosure of information, *Differential Privacy* (DP) (Dwork et al., 2006) has been integrated into machine learning (Abadi et al., 2016) and language models (McCann et al., 2017; Shi et al., 2022; Du et al., 2023). DP formalizes privacy through a notion of indistinguishability so that the model outputs are not affected by the addition or removal of an entry in the training corpus. This is accomplished by injecting additive noise on gradients during model training.

Due to scaling issues associated with DP on LMs during perturbation of per-sample gradient updates (Abadi et al., 2016), there is a trend towards perturbing the raw text (Fernandes et al., 2019; Feyisetan et al., 2020; Yue et al., 2021; Chen et al., 2023).

By exploiting the geometric proximity of words in word embeddings (Mikolov et al., 2013), Feyisetan et al. (2020) proposed a probabilistic mechanism grounded in metric DP (Chatzikokolakis et al., 2013) to perturb all words in a text while ensuring plausible deniability (Bindschaedler et al., 2017) of the text regarding its provenance and content.

However, several studies documented that mechanisms for embedding words in a high-dimensional space harbor (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Manzini et al., 2019) and transfer (Papakyriakopoulos et al., 2020) unwanted stereotypes and prejudices present in a text corpus.

Contribution. Building on the rich body of research exploring privacy-fairness trade-offs (Bagdasaryan et al., 2019; Farrand et al., 2020; Hansen et al., 2022), this study addresses the implications of text privatization on biased associations in LMs. Specifically, we pre-train BERT (Devlin et al., 2019) models with masked language modeling and next sentence prediction on webscraped text modified under varying levels of privacy. We then score the stereotypical bias following the context association test of Nadeem et al. (2021) and stereotype pairs benchmark of Nangia et al. (2020). Our findings reveal a nuanced landscape where stereotypical bias generally diminishes as privacy guarantees are tightened. This is in line with prior research indicating that LMs with impaired language modeling capabilities tend to exhibit less stereotypical associations (Nadeem et al., 2021). However, this diminution is not uniform across all social categories as biases associated with certain attributes show varying trends of stability, amplification, and attenuation. We thus advocate for careful bias measurement when deploying privacy-preserving LMs.

Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, pages 20–28
August 15, 2024 ©2024 Association for Computational Linguistics

2 Background

To ensure a consistent understanding of privacy and fairness in machine learning, we provide the foundations of differential privacy and a brief definition of stereotypical bias along with related work.

2.1 Differential Privacy

Differential Privacy (DP) (Dwork et al., 2006) originated in the field of statistical databases and was adapted to machine learning (Abadi et al., 2016). DP formalizes privacy through the indistinguishability of model outputs with respect to the presence or absence of a record in the dataset. The notion of indistinguishability is achieved through noise and can be controlled by the privacy budget $\varepsilon \in (0, \infty]$, with privacy guarantees diminishing as $\varepsilon \to \infty$.

Despite evidence of preventing information disclosure, the perturbations caused by noise can have detrimental (Jayaraman and Evans, 2019) and disparate (Bagdasaryan et al., 2019; Farrand et al., 2020; Hansen et al., 2022) effects on the behavior of machine learning models. By assessing the accuracy of differentially private machine learning models for (underrepresented) subgroups, Bagdasaryan et al. (2019) find a disparate impact regarding gender and ethnicity in both vision and text.

To prevent the risk of authorship disclosure, text rewriting is an appealing strategy that applies noise at word level or sentence level by leveraging word embeddings (Mikolov et al., 2013) or sequence-to-sequence models (Vaswani et al., 2017). Each approach comes with distinct mechanisms and implications for balancing utility and privacy.

Embedding-based Text Rewriting. Feyisetan et al. (2020) pioneered a mechanism for text rewriting termed `Madlib`. `Madlib` exploits the distance of words in embedding spaces (Mikolov et al., 2013) to substitute all words in a text with another word within a radius controlled by the privacy budget ε. Since this substitution mechanism scales the notion of indistinguishability by a distance, it satisfies the axioms of metric DP (Chatzikokolakis et al., 2013).

Building on a word embedding, the substitution involves three steps at word level: (1) retrieving the continuous representations of words from the embedding space, (2) adding noise to the representations calibrated using a multivariate distribution, and (3) mapping the noisy representation back onto the discrete space of vocabulary by employing a nearest neighbor approximation. While the probabilistic nature of these substitutions assures plausible deniability (Bindschaedler et al., 2017), substitutions based on the distance between words alleviate the curse of dimensionality typical of randomized response (Warner, 1965).

However, privatizing text through perturbations at word level imposes notable limitations. Since the privacy guarantee in this approach depend on the geometry of the embedding space, it necessitates meticulous calibration of the noise magnitude (Xu et al., 2020). For dense regions of the embedding space, excessive noise may obscure suitable substitutions. For sparse regions of the embedding space, minimal noise may not provide sufficient protection against reconstruction. In addition the to noise calibration, perturbations at word level, albeit retaining the meaning of a text, encounter difficulties in maintaining the coherence of the text, such as grammar (Mattern et al., 2022), ambiguity (Arnold et al., 2023), and hierarchy (Feyisetan et al., 2019).

Autoencoder-based Text Rewriting. Instead of privatization over word embeddings, an orthogonal approach utilizes sequence-to-sequence models built on recurrent (Bo et al., 2021; Krishna et al., 2021; Weggenmann et al., 2022) and transformer (Igamberdiev and Habernal, 2023) architectures. Common to these approaches is that noise is added to the encoder representations of text and the decoder learns to convert these noisy representations into text but without stylistic identifiers.

By perturbing the text at sentence level, this approach presents unique challenges compared to perturbing texts at word level. For instance, Igamberdiev et al. (2022) criticized that the utility is contingent upon the resemblance between the texts on which the sequence-to-sequence model was optimized and the texts that are subjected to privacy-preserving paraphrasing. This limitation in generalizability renders this form of text rewriting infeasible for the privatization of pretext at scale.

2.2 Stereotypical Bias

Bias in machine learning is viewed as prior information that informs algorithmic learning (Mitchell, 1980). When the prior information is predicated on stereotypes and prejudices, bias transcends this neutral definition and manifests in a disproportionate weight in favor of or against a social group.

The origins of these problematic biases are often rooted in the raw data used to develop machine learning models (Caliskan et al., 2017). Implicit or explicit stereotypes based on characteristics such as gender and race can cause the models to perpetuate and propagate these biases. This can significantly affect perception and decision making. The issue with stereotypical bias is particularly acute in the context of language models due to their extensive training on vast corpora that reflect biases present in human language. This bias magnifies the potential to influence its tone (Dhamala et al., 2021) and content (Abid et al., 2021), resulting in negative effects on individuals and society at large.

Using tests for association analogies, prior research demonstrated that embeddings harbor stereotypical biases related to gender (Bolukbasi et al., 2016; Kurita et al., 2019; Chaloner and Maldonado, 2019) and race (Manzini et al., 2019). Specifically, Caliskan et al. (2017) showed that terms related to career are associated with male names rather than female names, whereas unpleasant terms are associated with ethnic minorities. Garg et al. (2018) elaborate on the temporal dimension of bias in word embeddings by observing changes in gender and ethnic stereotypes over a century. This diachronic analysis indicates that while certain stereotypes have diminished over time, others remain robustly encoded in language. By investigating bias diffusion, Papakyriakopoulos et al. (2020) showed that biases contained in word embeddings can permeate natural language understanding, while Abid et al. (2021) report stereotypes in language generation such as violence for certain religious groups.

Unlike these studies on bias in raw data, we examine the bias that stems from text privatization.

3 Methodology

To test our hypothesis on amplification of stereotypical bias through text privatization, we need to define (1) a language model, (2) the mechanism for text privatization, and (3) a bias measurement.

3.1 Language Model

Following Qu et al. (2021), we use a BERT model (Devlin et al., 2019) leveraging masked language modeling and next sentence prediction tasks for pre-training. The choice of BERT is motivated by its widespread adoption and proven effectiveness in capturing contextual relationships within text.

For pre-training, we selected a webscraped replication of WebText (Radford et al., 2019), which compared to WikiText (Merity et al., 2016), covers a broader spectrum of topics, styles, and viewpoints. This diversity renders WebText particularly suited for examining the transfer of stereotypical biases from the pre-text corpus. For fine-tuning, we reproduced the experiments of Bagdasaryan et al. (2019) but found no stereotypical bias other than a disparate impact due to sampling bias.

To assess the alterations in stereotypical bias by text privatization, we trained a BERT model devoid of any privacy interventions, serving as a control to score amplification and attenuation, and three additional copies of the BERT model under varying degrees of privacy guarantees. Since all BERT models are identical in terms of architecture and optimization (differing solely in the degree of text privatization), this setup warrants a controlled comparison that isolates the effects of text privatization on the anchoring of stereotypical bias.

3.2 Text Privatization

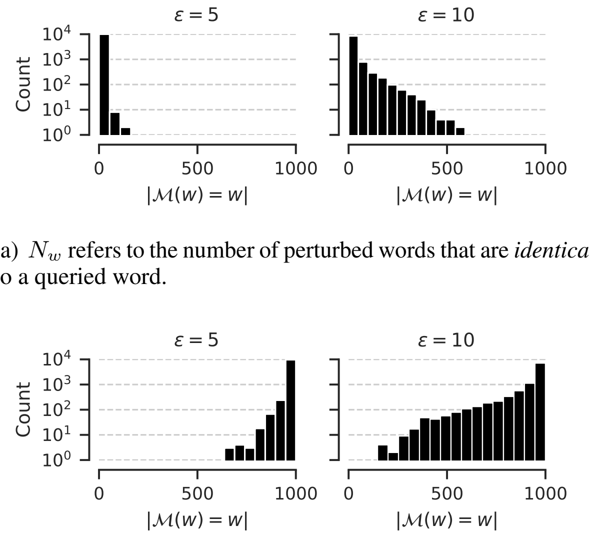
To privatize the WebText corpus, we operationalize the Madlib mechanism developed by Feyisetan et al. (2019) for text privatization at word level. Madlib necessitates the utilization of continuous representations supplied by a word embedding. We integrate Madlib with GloVe (Pennington et al., 2014). GloVe supplies a 400000-words vocabulary, each mapped to a 300-dimensional representation. The choice of GloVe is motivated by the richness of its semantic space, making it an ideal candidate for privacy-preserving text privatization.

Since the privacy guarantee of Madlib is rooted in metric DP, we need to calibrate the noise parameter ε according to the metric space of GloVe. This calibration involves an estimation of the plausible deniability (Bindschaedler et al., 2017) through two proxy statistics (Feyisetan et al., 2020):

- $N_w = \mathbb{P}\{M(w) = w\}$ measures the number of *identical* words that stem from perturbing a word given a privacy budget ε. We estimate N_w by counting the occurrence of unaltered words after querying a random subset of 10000 words for a total of 1000 times.

- $S_w = |\mathbb{P}\{M(w) = w'\}|$ measures the number of *unique* words that stem from perturbing a word given a privacy budget ε. We estimate S_w by calculating the effective support of a word after querying the same random subset of 10000 words for a total of 1000 times.

We can relate the proxy statistics to the privacy budget. Adding more noise corresponds to a tighter privacy guarantee. This is indicated by a smaller



(a) N_w refers to the number of perturbed words that are *identical* to a queried word.

(b) S_w refers to the number of perturbed words that are *unique* from a queried word.

Figure 1: Plausible deniability statistics approximated for a randomly compiled vocabulary of 10000 words, each word privatized over a number of 1000 queries.

value for ε and results in a diverse set of perturbed words (low N_w and high S_w). Adding less noise reflects a weaker privacy guarantee. This is characterized by a larger value for ε and results in more frequent unperturbed words (high N_w and low S_w).

Figure 1 presents the distribution of N_w and S_w. Since N_w (S_w) should be positively (negatively) skewed to assure a reasonable privacy guarantee, we adopt privacy budgets of $\varepsilon = \{5, 10\}$, corresponding to a high and low level of privacy protection, respectively. Table 1 illustrates an example obtained by querying `Madlib` using a privacy budget ε of 10. Notice the fidelity while some variation asserts compliance with privacy requirements.

3.3 Bias Measurement

Characterizing bias embedded within models typically relies on carefully crafted datasets. Several datasets exist to measure bias in word embeddings (Caliskan et al., 2017; May et al., 2019) and language models trained with masked (Nangia et al., 2020; Nadeem et al., 2021) and causal language modeling objective (Dhamala et al., 2021).

We adopt the `StereoSet` dataset designed by Nadeem et al. (2021). Given associative contexts, this dataset is intended to measure the tendency to default to stereotypical or anti-stereotypical associations. `StereoSet` provides meticulously crafted stimuli for bias measurement regarding gender, pro-

Table 1: Example sentence derived from `Webtext` and privatized for three independent runs of `Madlib` (Feyisetan et al., 2020) using a privacy budget ε of 10.

Tokens	**Substitutions**
Port-au-Prince	*rosita, xiangfan, tejgaon*
,	*and, as, ,*
Haiti	*vanuatu, cuba, haiti*
(	*(, 45, according*
CNN	*informed, journalist, speaker*
)	*–,), 2000*
–	*likely, –, two*
Earthquake	*quake, earthquake, stress*
victims	*killings, murdered, deaths*
,	*agrees, things, went*
writhing	*desolation, stayers ,tiredness*
in	*out, in, first*
pain	*frustration, fractures, pain*
and	*have, over, with*
grasping	*interplay, spit, dangling*
at	*at, the, as*
life	*proud, day, loves*
,	*and, took, 45*
watched	*watched, lined, raised*
doctors	*medical, researchers, surgeons*
and	*including, as, alongside*
nurses	*pharmacists, nurses, physicians*
walk	*walks, sideways, walked*
away	*gone, away, when*
from	*from, around, off*
a	*an, than, one*
field	*games, yards, field*
hospital	*school, nursing, staff*
Friday	*week, thursday, saturday*
night	*night, hours, watch*
after	*after, afterwards, before*
a	*a, first, one*
Belgian	*danish, macedonian, french*
medical	*medical, hospital, psychiatric*
team	*division, helm, cup*
evacuated	*evacuated, ferried, homeless*
the	*the, 1984, on*
area	*town, area, park*
,	*accused, 6, :*
saying	*asking, iranians, saying*
it	*since, as, is*
was	*that, only, subsequently*
concerned	*suspicious, expect, insist*
about	*nearly, just, about*
security	*beijing, actions, personnel*
.	*still, then, .*

fession, race, and religion at two distinct levels:

Intrasentence. The intrasentence task measures bias for sentence-level reasoning. It is formulated as a fill-mask task. Given a context sentence describing a social group, the task is to fill in a masked attribute corresponding to a stereotype, an anti-stereotype, and an unrelated option. The propensity for stereotypical associations is gauged by the likelihood of assigning each of these options.

Intersentence. The intersentence task measures bias for discourse-level reasoning. It is formulated as a next-sentence task. Given a context sentence pertaining to a social group, followed by three sentences embodying a stereotype, an anti-stereotype, and an unrelated attribute, the assessment of stereotypical bias hinges on which of these sentences is instantiated as the most likely continuation.

To capture social biases at more differentiated levels, we complement our investigation with the `CrowS-Pairs` benchmark designed by Nangia et al. (2020). This benchmark consists of pairs of minimally distant sentences dealing with bias about gender identity, ethnic affiliation, age, nationality, religion, sexual orientation, socioeconomic status, physical appearance, and disability. The first sentence in each pair demonstrates a stereotype about a social group, while the second sentence in each pair violates it. This allows to score the bias in a language model by measuring how frequently it prefers a statement that portrays a social group stereotypically compared to an alternative portrayal of the same situation with a different social identity.

Despite some criticism due to issues with model calibration (Desai and Durrett, 2020), we determine the preferences using pseudo-likelihood scoring (Salazar et al., 2020). We iterate over each sentence, masking a word at a time (except for the words that identify a social group), and accumulate the log-likelihoods of the masks in a sum for comparison.

4 Experiments

Prior to initiating our bias measurement, we conducted a preliminary sanity check by examining the pseudo-perplexity scores of `BERT` models trained under varying degrees of privacy. Pseudo-perplexity serves an indicator of a LM's ability to accurately model the probability distribution of words within a text corpus, thereby reflecting the model's proficiency to comprehend the linguistic structures encountered during its training.

Table 2: Percentage preference of stereotypical associations derived from `StereoSet`, where scores above 0.5 indicate pro-stereotypical bias and scores below 0.5 indicate anti-stereotypical bias. Effect sizes compared to the baseline value according to Cohens d in brackets.

Epsilon	∞	**10**	**5**
		Intrasentence	
Gender	**.6196**	.5490 (↓ .14)	.5020 (↓ .24)
Race	**.6060**	.5135 (↓ .19)	.4709 (↓ .27)
Religion	.5897	**.6538** (↑ .13)	**.6538** (↑ .13)
Profession	**.6062**	.5679 (↓ .08)	.5259 (↓ .16)
Average	**.6054**	.5711 (↓ .07)	.5382 (↓ .14)
		Intersentence	
Gender	.5868	**.5909** (↑ .01)	.5248 (↓ .12)
Race	.5318	.5287 (↓ .01)	**.5461** (↑ .03)
Religion	**.5641**	.5513 (↓ .03)	.5385 (↓ .05)
Profession	**.6070**	.5272 (↓ .16)	.4813 (↓ .25)
Average	**.5724**	.5495 (↓ .05)	.5227 (↓ .10)

We use a 10% subset of `WikiText` for computing the pseudo-perplexities. Evaluated at privacy levels specified by the privacy parameter ε, the pseudo-perplexity scores were 93.51 with no privacy interventions, 502.67 with moderate privacy settings, and 2056.43 under conditions of high privacy. Consistent with previous evidence that introducing noise at word-level compromises the linguistic proficiency of LMs (Mattern et al., 2022), these results demonstrate a substantial degradation as the level of privacy augmentation increases.

The observed degradation raises an interesting question of whether private LMs harbor stereotypical biases despite diminished language modeling capabilities. This question forms the basis for our subsequent analysis of the undesirable biases in LMs stemming from text privatization.

4.1 Stereotype Results from StereoSet

To measure the bias resulting from text privatization at sentence and discourse level, we commence our analysis by detailing the stereotype scores derived from the `StereoSet` benchmark. The stereotype score is defined by the percentage of examples for which the LM assigns a higher probability to the pro-stereotypical word as opposed to the anti-stereotypical word. As such, scores closer to 0.5 are indicative of unbiased associations.

Table 2 presents the averaged stereotype scores grouped by intrasentence and intersentence tasks

Table 3: Percentage preference of stereotypes derived from `CrowS-Pairs`, where scores closer to 0.5 are indicative of unbiased associations. Effect sizes of text privatization compared to the baseline value in brackets.

Epsilon	∞	**10**	**5**
Gender	.5229	**.5878** (↑ .13)	.5267 (↑ .01)
Age	.4943	.4943 (↑ .00)	**.5402** (↑ .09)
Race	.5233	.5446 (↑ .04)	**.5640** (↑ .08)
Religion	**.6000**	.5905 (↓ .02)	.5905 (↓ .02)
Nationality	.5283	**.5535** (↑ .05)	.5346 (↑ .01)
Occupation	**.5465**	.5407 (↓ .01)	.4535 (↓ .19)
Sexuality	**.6786**	.6190 (↓ .12)	.5119 (↓ .34)
Disability	**.6167**	.6000 (↓ .03)	.5500 (↓ .13)
Appearance	.4762	**.6190** (↑ .29)	.4921 (↑ .03)

and segmented by social categories [1]. Several key trends inform our understanding of the impact of text privatization on stereotypical bias. We observe that results from the intrasentence task aligns with those from the intersentence task, showing that the stereotype scores decline as the privacy level intensifies. For the intrasentence tasks, the averaged stereotype scores decreased from 0.6054 to 0.5711 and 0.5382 as the privacy budget was tightened to 10 and 5, respectively. For the intersentence tasks, the stereotype scores decreased similarity from 0.5724 to 0.5495 and 0.5227, respectively. However, the fall in stereotype scores is overall more pronounced in the intrasentence task than in the intersentence task. This disparity implies that mask language modeling is affected more acutely than next sentence prediction, which requires a broader context to build stereotypical association.

While text privatization generally reduces stereotypical biases, we find inconsistent pattern when breaking down the stereotype scores by social categories. This indicates that the impact of text privatization is not uniformly spread across social groups.

4.2 Stereotype Results from CrowS-Pairs

To explore the manifestation of stereotypical bias across a broader range of social categories, we broadened our analysis to include `CrowS-Pairs`. Table 3 confirms that there is no overarching trend regarding the degree of text privatization and the manifestation of stereotypical biases.

Following the general observation of decreasing stereotype scores as the privacy budget tightens, further scrutiny into social categories reveals a complex and heterogeneous response to text privatization. We discern social categories that are constant (e.g., religion), amplified (e.g., age, race), and attenuated (e.g., occupation, sexuality, disability). This suggests that some social categories are detached from the influences of textual perturbations while others seem less robust. Further complicating the interactions is that some social categories (e.g., gender, nationality, appearance) experience fluctuating responses. The categories show an increase in stereotype scores as privacy settings are intensified before stabilizing or reverting at the strictest levels of privacy. Except for sexual orientation (↓ .34) and physical appearance appearance (↑ .29), the effect sizes are negligible. This variability underscores the intricate dynamics between text privatization and LMs, suggesting that minor modifications in the privacy parameters can have significant and diverse impacts on stereotypical biases across different social constructs.

5 Conclusion

The interaction dynamics that govern the manifestation of bias in LMs are equivocal (Hansen et al., 2022). Prior research indicates that stereotypical bias is related to language proficiency in LMs (Nadeem et al., 2021). Since text privatization is known to impair language modeling capabilities (Feyisetan et al., 2020), one would expect a general diminution of stereotypical bias. However, the word embeddings used for text privatization are documented to harbor (Bolukbasi et al., 2016; Caliskan et al., 2017) and transfer (Papakyriakopoulos et al., 2020) stereotypical biases. This duality raises questions about whether text privatization leads to an amplification or an attenuation of stereotypical biases. By probing a LMs tendency to default to stereotypical or anti-stereotypical associations, we aimed to elucidate the relationship between text privatization and the amplification or attenuation of biases. We find that different social domains react differently to privacy settings and recommend to carefully assess stereotypical bias after training a LM on a privatized corpus of text.

[1] Since `Madlib` involves a probabilistic mechanisms, one could argue that the bias patterns of the privacy budget ε on social categories is caused by the randomness of text privatization. To test whether the observed patterns stem from randomness, we reproduced all experiments using three distinct seeds. The variance across different configurations suggests that these patterns are inherent to the privatization process and not merely artifacts of random perturbations.

6 Limitations

This study has several limitations that warrant consideration. Our experiments are based on WebText. While this corpus provides a broad range of topics and styles, it is possible that the derived insights, such as the general reduction in stereotypical bias and the unequal reduction across social groups, are influenced by spurious correlations (Schwartz and Stanovsky, 2022) inherent in the dataset. In addition to the flaws caused by the training corpus, our reliance on GloVe embeddings for text privatization introduces another potential source of inherent biases. Future research should address these limitations by incorporating a more diverse set of datasets and explore how alternative embeddings affect the persistence of stereotypical bias after privatization.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Driving context into text-to-text privatization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 15–25, Toronto, Canada. Association for Computational Linguistics.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*.

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3997–4007, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2665–2679.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, pages 15–19.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Victor Petren Bach Hansen, Atula Tejaswi Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Sogaard. 2022. The impact of differential privacy on group disparity mitigation. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 12–12, Seattle, United States. Association for Computational Linguistics.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tom M Mitchell. 1980. The need for biases in learning generalizations.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

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

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 446–457.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1488–1497.

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

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Roy Schwartz and Gabriel Stanovsky. 2022. On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390.

Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *International Conference on Text, Speech, and Dialogue*, pages 273–281. Springer.

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

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.