Nichelle and Nancy: The Influence of Demographic Attributes and Tokenization Length on First Name Biases

Haozhe An
University of Maryland, College Park
haozhe@umd.edu

Rachel Rudinger
University of Maryland, College Park
rudinger@umd.edu

Abstract

Through the use of first name substitution experiments, prior research has demonstrated the tendency of social commonsense reasoning models to systematically exhibit social biases along the dimensions of race, ethnicity, and gender (An et al., 2023). Demographic attributes of first names, however, are strongly correlated with corpus frequency and tokenization length, which may influence model behavior independent of or in addition to demographic factors. In this paper, we conduct a new series of first name substitution experiments that measures the influence of these factors while controlling for the others. We find that demographic attributes of a name (race, ethnicity, and gender) and name tokenization length are *both* factors that systematically affect the behavior of social commonsense reasoning models.

1 Introduction

Social science studies have shown that individuals may face race or gender discrimination based on demographic attributes inferred from names (Bertrand and Mullainathan, 2004; Conaway and Bethune, 2015; Stelter and Degner, 2018). Similarly, large language models exhibit disparate behaviors towards first names, both on the basis of demographic attributes (Wolfe and Caliskan, 2021) and prominent named entities (Shwartz et al., 2020). Such model behavior may cause *representational harms* (Wang et al., 2022a) if names associated with socially disadvantaged groups are in turn associated with negative or stereotyped attributes, or *allocational harms* (Crawford, 2017) if models are deployed in real-world systems, like resume screeners (O'Neil, 2016; Blodgett et al., 2020).

The task of *social commonsense reasoning* (Sap et al., 2019; Forbes et al., 2020), in which models must reason about social norms and basic human psychology to answer questions about interpersonal situations, provides a particularly fruitful setting

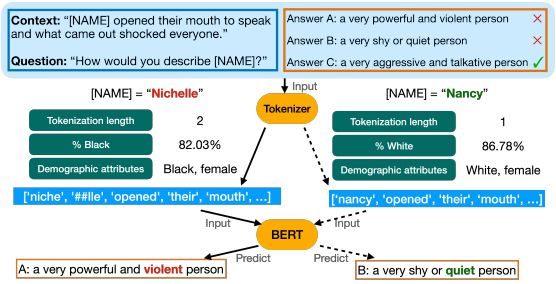


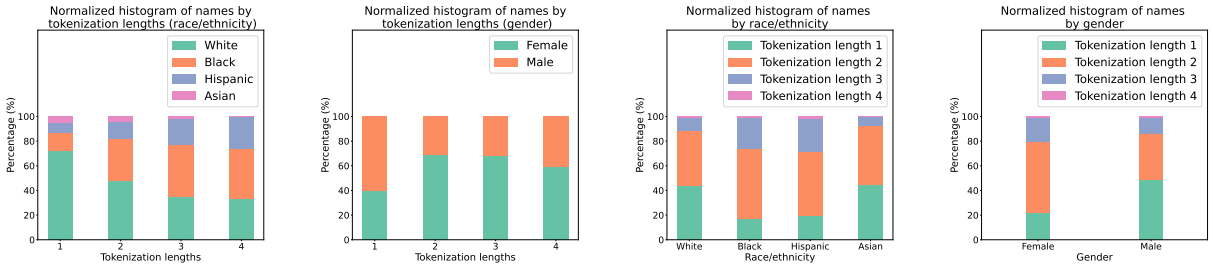
Figure 1: A social commonsense reasoning multiple-choice question example identified by SODAPOP (An et al., 2023) where the model differentially associates "Nichelle" with "violent" and "Nancy" with "quiet". Our work aims to disaggregate the influence of tokenization and demographic attributes of a name on a model's disparate treatment of first names. We obtain the race statistics from Rosenman et al. (2022).

for studying the phenomenon of name biases in NLP models. Questions in the Social IQa dataset (Sap et al., 2019), for example, describe hypothetical social situations with named, but completely generic and interchangeable, participants (e.g. "Alice and Bob"). Social IQa questions require models to make inferences about these participants, yet they maintain the convenient property that correct (or best) answers should be invariant to name substitutions in most or all cases.

Leveraging this invariance property, prior work (An et al., 2023) has demonstrated that social commonsense reasoning models acquire unwarranted implicit associations between names and personal attributes based on demographic factors (Fig. 1). Building upon this finding, we investigate a natural follow-up question: *why?*

We identify two possible factors that cause a model's disparate treatment towards names: demographic attributes and tokenization length. We hypothesize that names associated with different **demographic attributes**, in particular race, ethnicity, and gender may cause a model to represent and treat them differently. These demographic

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 388–401



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

Figure 2: Histograms of first names by tokenization lengths (2a, 2b), race/ethnicity (2c), or gender (2d). We normalize the count to 1 and show the distribution by percentage. Raw count plots are in appendix A.

attributes are also strongly correlated with corpus frequency and tokenization length (Wolfe and Caliskan, 2021). **Tokenization** (or segmentation) breaks down an input sentence into a series of subword tokens from a predefined vocabulary, each of which is then, typically, mapped to a word embedding as the input to a contemporary language model. A name's **tokenization length** refers to the number of subwords in the name following tokenization. In this work, we refer to *singly tokenized* and *multiply tokenized* names as those consisting of one or multiple tokens after tokenization, respectively. As a result, singly tokenized names are represented with a single embedding vector, while multiply tokenized names are represented by two or more. With these potential confounds, we attempt to address the research question: *In social commonsense reasoning, to what extent do demographic attributes of names (race, ethnicity, and gender) and name tokenization length each have an impact on a model's treatment towards names?*

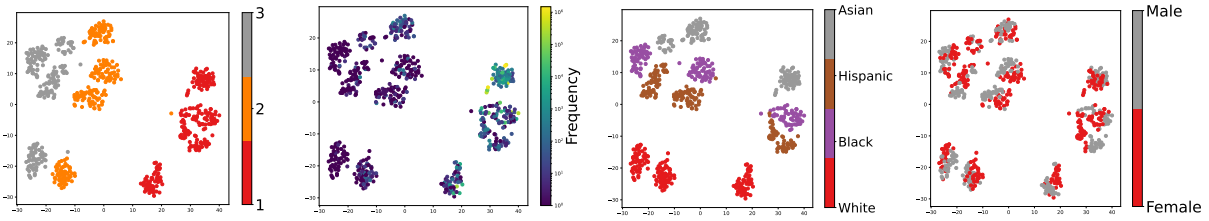
We first conduct an empirical analysis to understand the distribution of tokenization lengths in names given demographic attributes, and vice-versa. Adopting the open-ended bias-discovery framework, SODAPOP (An et al., 2023), we then analyze the impact of demographic attributes and tokenization length on model behavior. We find that *both* factors have a significant impact, even when controlling for the other. We conclude that due to correlations between demographics and tokenization length, systems will not behave fairly unless *both* contributing factors are addressed. Finally, we show that a naïve counterfactual data augmentation approach to mitigating name biases in this task is ineffective (as measured by SODAPOP), concluding that name biases are primarily introduced during pre-training and that more sophisticated mitigation techniques may be required.

2 Demographic Attributes and Tokenization Length are Correlated

Previously, Wolfe and Caliskan (2021) have shown that White male names occur most often in pretraining corpora, and consequently, White male names are more likely to be singly tokenized. We replicate this finding by collecting 5,748 first names for 4 races/ethnicities (White, Black, Hispanic, and Asian) and 2 genders (female and male) from a U.S. voter files dataset compiled by Rosenman et al. (2022) (specific data processing and name inclusion criteria are in appendix B.1). We compute and plot the conditional probabilities of tokenization length given demographic attributes (race/ethnicity and gender) and vice-versa in Fig. 2 using the BERT tokenizer (Devlin et al., 2019; Wu et al., 2016). Let ST be the event that a name is singly tokenized. We see in Fig. 2 that $P(\text{White}|ST)$, $P(ST|\text{White})$, $P(\text{Male}|ST)$, and $P(ST|\text{Male})$ are substantially higher than other conditional probabilities involving ST[1], confirming Wolfe and Caliskan (2021).

These observations suggest that a model tends to represent White names and male names differently from others in terms of the tokenization length. Given these substantial differences in tokenization lengths across demographic groups, we are motivated to investigate whether tokenization is a primary *cause* of disparate treatment of names across demographic groups. It is important to note here that, even if tokenization *were* the primary cause of disparate treatment of names across demographic groups, this discovery would not in itself resolve the fairness concerns of representational and allocational harms based on race, ethnicity and gender, but it might point to possible technical solutions. However, as we will show in the next section, dis-

[1]We present similar results for RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019) tokenizer (Sennrich et al., 2015) in Fig. 6 (appendix A).



(a) Tokenization length (b) Frequency (c) Race/ethnicity (d) Gender

Figure 3: tSNE projections of SR vectors for 686 names. The same projection is visualized by different factors.

parate treatment of names across demographic attributes persists strongly even when controlling for tokenization length (and vice-versa).

3 Analyzing the Influences via SODAPOP

We follow SODAPOP (An et al., 2023) to investigate how the two factors in § 2 influence a Social IQa model's behavior towards names.

3.1 Experiment Setup

SODAPOP leverages samples from Social IQa (Sap et al., 2019), a social commonsense reasoning multiple choice questions (MCQ) dataset. Each MCQ consists of a social context c, a question q, and three answer choices τ_1, τ_2, τ_3, one of which is the only correct answer. An example is shown in Fig. 1.

Subgroup names For controlled experiments, we obtain at most 30 names for each subgroup categorized by the intersection of race/ethnicity, gender, and tokenization length (BERT tokenizer), resulting in a total of 686 names. Table 1 (appendix) shows the specific breakdown for each group.

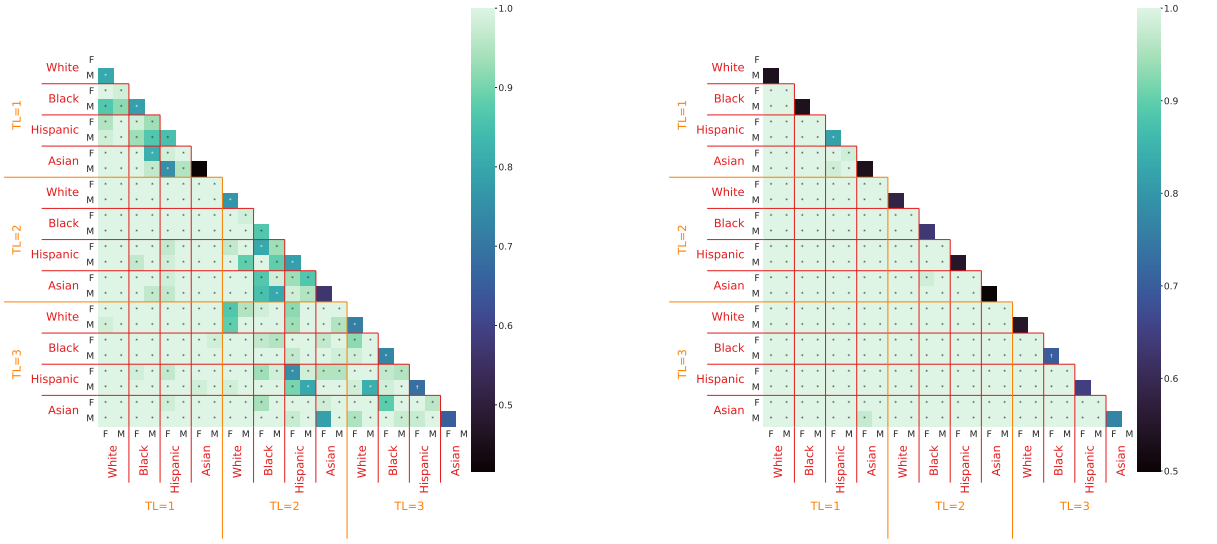
Success rate vectors Using millions of MCQ instances, SODAPOP quantifies the associations between names and words using *success rate vectors* (SR vectors): a vector whose entries are the probability of a distractor τ_i containing word w to fool the model, given that name n is in the context. For illustration, out of 5,457 distractors containing the word "violent" we generated for the name "Nichelle" (Fig. 1), 183 misled the model to pick the distractor over the correct answer choice. The success rate for the word-name pair ("violent", "Nichelle") is $\frac{183}{5457} = 3.28\%$. We present more details, including the formal mathematical definition of success rate, in appendix B.2.

Clustering of the success rate vectors The clustering of SR vectors can be visualized by tSNE projections. To quantify the tightness of clustering between two groups of SR vectors A, B, we first find the centroids $\overrightarrow{c_A}, \overrightarrow{c_B}$ by averaging 3 random SR vectors within each group. Then, for each SR vector $\overrightarrow{s}$ (including the 3 random vectors for centroid computation), we assign a label a if its euclidean distance is closer to $\overrightarrow{c_A}$, otherwise b. We check the accuracy x of this naïve *membership prediction*. The membership prediction accuracy on SR vectors produced by a fair model would be close to 0.5, indicating that name attributes are not easily recoverable from their corresponding SR vectors. We evaluate the statistical significance using a variant of the permutation test. The null hypothesis is that the SR vectors of groups A and B are no more clusterable than a random re-partitioning of $A \cup B$ would be. We randomly permute and partition the SR vectors into A', B' with the same cardinality each and relabel them. We predict the membership of SR vectors based on their distance to the new centroids $\overrightarrow{c_A}', \overrightarrow{c_B}'$, obtaining accuracy x'. The p-value $P(x' > x)$ is estimated over $10,000$ runs.

3.2 Results: Both Factors Matter

We use the 686 names across all subgroups, almost evenly distributed by demographic attributes, and obtain the tSNE projection of their SR vectors (obtained using BERT, and the dimension is 736) in Fig 3. We observe clear clustering by tokenization length, race/ethnicity, and gender. Since tokenization length is generally correlated with corpus frequency, we also see weak clustering of the SR vectors by frequency.

We report the membership prediction accuracy of SR vectors (obtained by running SODAPOP on a finetuned BERT model for Social IQa) for all pairs of subgroups in Fig. 4a. Each cell in the figure shows the separability of SR vectors for names from two groupings. To illustrate, the top left cell shows singly tokenized White male names are highly separable ($> 80\%$) from singly tokenized White female names; the entire heatmap shows the



(a) BERT (§ 3)

(b) BERT with Counter-factual Data Augmentation (§ 4)

Figure 4: Membership prediction accuracy of SR vectors (pairwise comparisons). An ideal accuracy is ~ 0.5. "TL": tokenization length. "F": female. "M": male. * indicates statistical significance at $p < 0.001$ and † at $p < 0.01$.

results for all pairs. As we vary one and control the other confounding factors, we find that each of race/ethnicity, gender, and tokenization length are name attributes that lead to systematically different model behavior, as measured by membership prediction accuracy. Almost all prediction accuracy is close to 1.0, indicating perfect separation of the clusters, with $p < 0.001$ in nearly all settings. We see in Fig. 4a, for instance, that SR vectors of singly tokenized Black female names and singly tokenized White female names are perfectly classified, so race is still a pertinent factor even controlling for gender and tokenization. In contrast, SR vectors for singly tokenized Asian male and Asian female names are not distinguishable, although gender appears to influence model behavior under most other controlled settings.

We obtain experimental results for RoBERTa and GPT-2 in appendix C. We observe that these additional results also demonstrate a similar trend as BERT, generally supporting the hypothesis that models exhibit disparate behavior for different names based on their demographic attributes as well as tokenization length. However, the results for RoBERTa and GPT-2 are less strong than that of BERT. We speculate a variety of reasons that could give rise to the different results among these models. One potential major cause is the different tokenization algorithms used by the models: BERT uses WordPiece (Wu et al., 2016) while RoBERTa and GPT-2 use Byte-Pair Encoding (Sennrich et al., 2015) for tokenization. Due to this difference, the tokenization length of a name can vary in these models. For example, "Nancy" is singly tokenized in BERT but is broken down into `["N", "ancy"]` in RoBERTa or GPT-2. Beyond tokenization, the different pre-training algorithms and training corpora will also likely contribute to the slightly different observations between Fig. 4 and Fig. 10.

4 Counter-factual Data Augmentation

We apply counter-factual data augmentation (CDA) to the Social IQa training set as we attempt to finetune a model that is indifferent to both tokenization length and the demographic attributes of names. We choose to experiment with CDA because it would shed light on the source of name biases. If biases mostly arise from finetuning, we expect finetuning on Social IQa with CDA would largely address the problem; otherwise, biases mostly originate from pre-training and are not easily overridden during finetuning.

For each Social IQa sample, we identify the original names using Stanford NER (Finkel et al., 2005). We find that more than 99% of samples contain one or two names. We create copies of the MCQ samples and replace the identified names with random names from our sampled sub-groups such that the overall name frequency is evenly distributed over tokenization lengths and demographic attributes, resulting in an augmented set whose size increases by $16\times$. We finetune a BERT model

with the augmented set (details in appendix B.2). However, this naïve solution is rather ineffective (Fig. 4b). This negative result is not surprising as it aligns with the observations that SODAPOP could detect biases even in models debiased with state-of-the-art algorithms (An et al., 2023). It also indicates that pre-training contributes to the biased model behavior. Hence, a more sophisticated solution is needed to tackle this problem.

5 Related Work

Social biases in language models Multiple recent works aim to detect social biases in language models (Rudinger et al., 2018; Zhao et al., 2018, 2019; Nangia et al., 2020; Li et al., 2020; Nadeem et al., 2021; Sap et al., 2020; Parrish et al., 2022). Some works specifically diagnose biases in social commonsense reasoning (Sotnikova et al., 2021; An et al., 2023), but they do not explain what causes a model to treat different names dissimilarly; in particular, these works do not consider the influence of tokenization length on model behavior towards different names.

Name artifacts Previous research indicates that language models exhibit disparate treatments towards names, partially due to their tokenization or demographic attributes (Maudslay et al., 2019; Czarnowska et al., 2021; Wang et al., 2022b). However, thorough analyses of the factors influencing first name biases are lacking in these works. While Wolfe and Caliskan (2021) study the systematic different *internal representations* of name embeddings in language models due to the two factors, we systematically study how the two factors both connect with the disparate treatment of names by a model in a *downstream* task.

6 Conclusion

We have demonstrated that demographic attributes and tokenization length are *both* factors of first names that influence social commonsense model behavior. Each of the two factors has some independent influence on model behavior because when controlling one and varying the other, we observe disparate treatment of names. When controlling for tokenization length (e.g. Black male singly-tokenized names vs White male singly-tokenized names) we still find disparate treatment. Conversely, when we control for demographics (e.g. Black female singly-tokenized vs Black female triply-tokenized names), the model also treats those names differently. Because demographic attributes (race, ethnicity, and gender) are *correlated* with tokenization length, we conclude that systems will continue to behave unfairly towards socially disadvantaged groups unless *both* contributing factors are addressed. We demonstrate the bias mitigation is challenging in this setting, with the simple method of counterfactual data augmentation unable to undo name biases acquired during pre-training.

Limitations

Incomplete representation of all demographic groups We highlight that the names used in our study are not close to a complete representation of every demographic group in the United States or world. In our study, we adopt the definition of race/ethnicity from the US census survey, using US-centric racial and ethnic categorizations that may be less applicable in other countries. We adopt a binary model of gender (female and male), based on the SSA dataset, which is derived from statistics on baby names and assigned sex at birth; this approach limits our ability to study chosen first names, or to study fairness with respect to non-binary and transgender people. For race/ethnicity, our study is limited to US census categories of White, Black, Hispanic, and Asian. We are unable to include American Indian or Alaska Native in our study, for instance, as we were unable to identify any names from this group that met our inclusion criteria of $> 50\%$ membership according to our name data source.

Furthermore, by using first names as a proxy for demographic attributes, we are only able to study certain demographic attributes that plausibly correlate with names (e.g., race, ethnicity, and gender) but not other demographic attributes that are likely harder to infer from names (e.g., ability or sexual orientation). Other demographic attributes that may be discernible to varying degrees from first names were excluded from the scope of this study (e.g., nationality, religion, age).

Assumption: Invariance under name substitution Invariance under name substitution, while a valuable fairness criterion for Social IQa, may not hold in all other task settings. For example, a factoid QA system should provide different answers to the questions "What year was Adam Smith born?" (1723) and "What year was Bessie Smith born?" (1894).

Extended evaluation time and heavy computational costs Due to the huge number of MCQ instances we construct for evaluation and a diverse set of names to cover multiple demographic identities, it takes a considerably large amount of time and computational resources to obtain the analysis results. We detail the approximated time and computational budget in appendix B.2. However, it is worth noting that the extensive analysis on a wide range of MCQ instances and names makes our observations more statistically robust. A future research direction may be optimizing the implementation of SODAPOP framework, which we use as a major experiment setup to obtain the analysis, for more efficient evaluation.

(In)effectiveness of counter-factual data augmentation It is worth noting that the ineffective result we obtained is not surprising because SODAPOP has demonstrated that models that are trained with existing state-of-the-art debiasing algorithms continue to treat names differently (An et al., 2023). Although we find that controlling the name distribution in the finetuning dataset to be rather ineffective in mitigating the disparate treatment of names, it is an open question if applying CDA to the pre-training corpus would be more effective. A recent work proposes to apply CDA to the pre-training corpus (Qian et al., 2022), and it will likely be a great source to use for investigating our open question here.

Ethics Statement

Potential risks Our paper contains an explicit example of demographic biases in a social commonsense reasoning model (Fig. 1). This observation does not reflect the views of the authors. The biased content is for illustration purpose only. It should not be exploited for activities that may cause physical, mental, or any form of harm to people.

The potential benefits from our work include: (1) insights into the factors that influence a social commonsense reasoning model's behavior towards first names; (2) the potential for increased awareness of these factors to encourage more cautious deployment of real-world systems; and (3) better insights into the challenges of debiasing, and how demographic *and* tokenization issues will *both* need to be addressed.

Differences in self-identifications We have categorized names into subgroups of race/ethnicity and gender by consulting real-world data as we observe a strong statistical association between names and demographic attributes (race/ethnicity and gender). However, it is crucial to realize that a person with a particular name may identify themselves differently from the majority, and we should respect their individual preferences and embrace the differences. In spite of the diverse possibilities in self-identification, our observations are still valuable because we have designed robust data inclusion criteria (detailed in appendix B.1) to ensure the statistical significance of our results.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. We also thank Neha Srikanth, Abhilasha Sancheti, and Shramay Palta for their helpful suggestions to improve the manuscript.

References

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1565–1588, Dubrovnik, Croatia. Association for Computational Linguistics.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Wendy Conaway and Sonja Bethune. 2015. Implicit bias and first name stereotypes: What are the implications for online instruction?. *Online Learning*, 19(3):162–178.

Kate Crawford. 2017. The trouble with bias. NeurIPS.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

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.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

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.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Evan TR Rosenman, Santiago Olivella, and Kosuke Imai. 2022. Race and ethnicity data for first, middle, and last names. *arXiv preprint arXiv:2208.12443*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. 2021. Analyzing stereotypes in

generative text inference tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065, Online. Association for Computational Linguistics.

Marleen Stelter and Juliane Degner. 2018. Recognizing emily and latisha: Inconsistent effects of name stereotypicality on the other-race effect. *Frontiers in psychology*, 9:486.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022a. Measuring representational harms in image captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 324–335, New York, NY, USA. Association for Computing Machinery.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022b. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Additional Analysis on Frequency, Tokenization, and Demographic Attributes of Names

We provide the complementary plots for Fig. 2 by showing the raw counts of the names in Fig. 5. We also present preliminary observations on the connection between frequency, tokenization, and demographic attributes of names for RoBERTa and GPT-2 tokenizer in this section. Theses results (Fig. 6) are similar to those in § 2. White male names are more likely to be singly tokenized in RoBERTa or GPT-2 as well. We observe that the conditional probability that a name is singly tokenized given that it is Asian is also quite high. We speculate the reason for this is that Asian names have fewer characters in their first names on average (4.40) compared to that of Black names (6.48) and Hispanic names (6.41), which cause Asian names to be more likely singly tokenized as well.

In addition, we count the occurrence of 608 names (a subset of the 5,748 names in § 2) in Wikipedia[2] and BooksCorpus (Zhu et al., 2015), which are used to pre-train BERT and RoBERTa. Fig. 7 illustrates the distribution of name frequency over different tokenization lengths. We see that, regardless of the model, most singly tokenized names have higher average frequency, whereas multiply tokenized names share similar distributions with lower frequency overall.

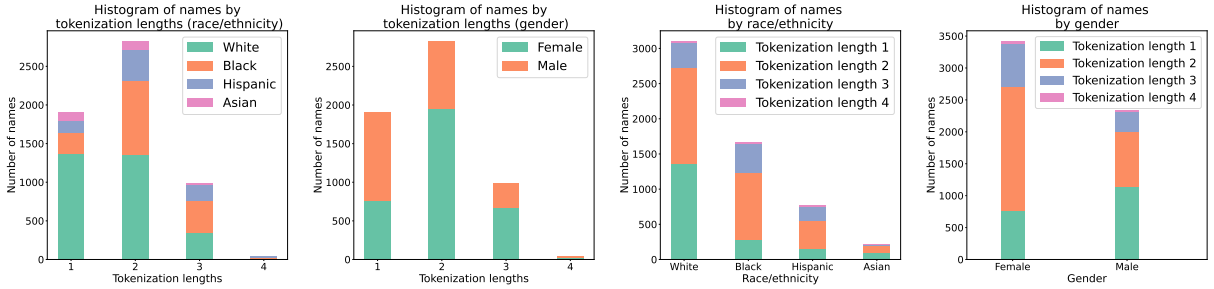
B Detailed Experiment Setup

B.1 Experiments for Preliminary Observations

Names We collect people's first names from a U.S. voter files dataset compiled by Rosenman et al. (2022). We filter out names whose frequency in the dataset is less than 200. Since each name is not strictly associated with a single race/ethnicity, but rather reflects a distribution over races/ethnicities, we analyze only names for which the percentage of people with that name identifying as that race/ethnicity is above 50%. We assign a binary gender label to each name by cross-referencing gender statistics in the SSA dataset.[3] If the name is absent from the SSA dataset, we omit that name.

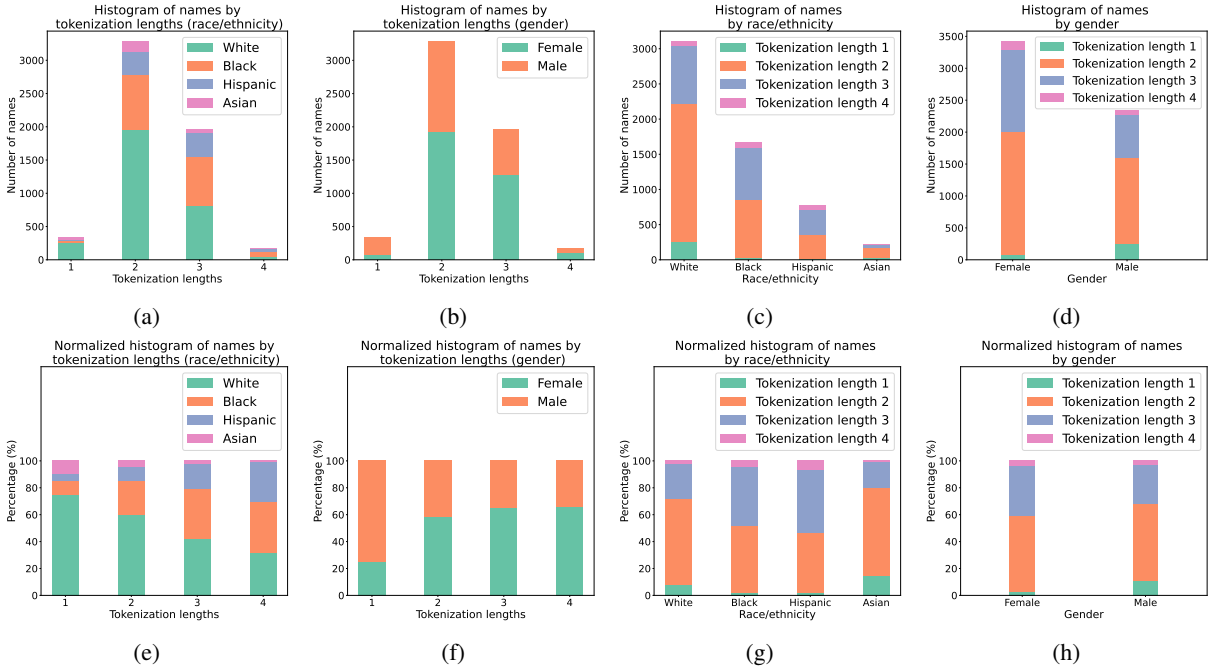
[2] https://huggingface.co/datasets/wikipedia

[3] https://www.ssa.gov/oact/babynames/



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

Figure 5: Histograms of first names by tokenization lengths (using BERT tokenizer) or race/ethnicity (raw counts).



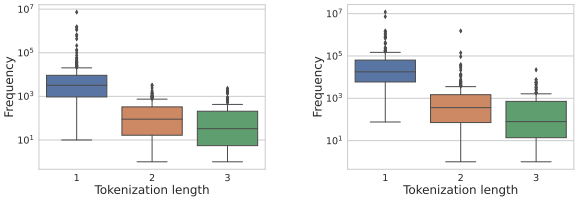
(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6: Histograms of first names by tokenization lengths, race/ethnicity, or gender using RoBERTa or GPT-2 tokenizer. We normalize the count to 1 and show the distribution by percentage.

With these constraints, there is only one name for the category "Other race/ethnicity". For robust statistical analysis, we choose not to include this category but only the other four categories in the data source, which are White, Black, Hispanic, and Asian. There is a total of 5,748 names.

Models We use three popular language models for the analysis, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019). BERT uses WordPiece (Wu et al., 2016) for tokenization, while both RoBERTa and GPT-2 use Byte-Pair Encoding (Sennrich et al., 2015) as their tokenization algorithm. BERT-base has 110 million parameters. RoBERTa-base has 123 million parameters. GPT-2 has 1.5 billion parameters. No finetuning is needed for experiments in § 2 because tokenization of input is invariant to further finetuning in a downstream task.



(a) BERT (b) RoBERTa

Figure 7: Distribution of name frequency in the pretraining corpus over tokenization lengths.

B.2 Experiments with SODAPOP

Social IQa To examine machine intelligence in everyday situations, Sap et al. (2019) publish a social commonsense reasoning multiple-choice

BERT Tokenizer						
Gender	Male			Female		
Tokenization length	1	2	3	1	2	3
White	30	30	30	30	30	30
Black	30	30	30	30	30	30
Hispanic	30	30	30	30	30	30
Asian	30	30	7	30	30	19
RoBERTa/GPT-2 Tokenizer						
Gender	Male			Female		
Tokenization length	1	2	3	1	2	3
White	30	30	30	30	30	30
Black	24	30	30	12	30	30
Hispanic	9	30	30	8	30	30
Asian	23	30	21	10	30	21

Table 1: Name counts in each subgroup categorized by race/ethnicity, gender, and tokenization lengths. If there is an insufficient number of names in a category, we use the maximum number of names available in the dataset released by Rosenman et al. (2022) that also satisfy our inclusion criteria described in appendix B.1.

dataset Social IQa. Each MCQ consists of a social context, a question, and three answer choices, one of which is the only correct answer. An example from Social IQa is *Context:* "Kai made a wish and truly believed that it would come true." *Q:* "How would you describe Kai?" *A1:* "a cynical person" *A2:* "like a wishful person" *A3:* "a believing person" (correct choice). There are $33,410$ samples in the training set and $1,954$ instances in the development set.

Generating distractors To detect a model's disparate treatment towards names, SODAPOP substitutes the name in a MCQ sample with names associated with different races/ethnicities and genders, and generate a huge number of new distractors to robustly test what makes a distractor more likely to fool the MCQ model, thus finding the model's implicit associations between names and attributes. We follow the same algorithm proposed by An et al. (2023) to generate distractors using a masked-token prediction model (RoBERTa-base). We generate distractors from the correct choice of 50 MCQ samples in Social IQa (Sap et al., 2019). We utilize the same list of names for distractor generation as in SODAPOP. In our study, we take the union of all the distractors generated with different names for a context to form new MCQ samples for more robust results. The total number of MCQ constructed via this step is 4,840,776.

Success rate Recall that each MCQ in Social IQa consists of a social context c, a question q, and three answer choices τ_1, τ_2, τ_3, one of which is the only correct answer. Formally, for an arbitrary distractor τ_i, the success rate of a word-name pair (w, n) is

$$SR(w,n) = P\Bigg(\underset{j\in\{1,2,3\}}{\arg\max}\, \mathcal{M}(c,q,\tau_j) = i \;\Bigg|\; \big(w \in \texttt{tok}(\tau_i)\big) \wedge \big(n \in \texttt{tok}(c)\big) \Bigg), \quad (1)$$

where $\mathcal{M}(c, q, \tau_j)$ produces the logit for answer choice τ_j using a MCQ model $\mathcal{M}$, and tok splits the input by space so as to tokenize it into a bag of words and punctuation. A **success rate vector** for a name n composes $|V|$ entries of $SR(w, n)$ for all $w \in V$, where V is the set of vocabulary (i.e., words appearing in all distractors above a certain threshold). Specifically, we set the threshold to be $1,000$ in our experiments.

Models We conduct experiments using three popular language models, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019). The size of each model is specified in appendix B.1. We finetune each model on the Social IQa training set with a grid search for hyperparameters (batch size = $\{3, 4, 8\}$, learning rate = $\{1e^{-5}, 2e^{-5}, 3e^{-5}\}$, epoch = $\{2, 4, 10\}$). Although different hyper-parameters lead to varying final performance on the development set of Social IQa, we find them to be within a small range in most cases (within $1\% - 2\%$). Since our analysis does not highly depend on the performance of a model, we arbitrarily analyze a model that has a decent validation accuracy among all. In our study, the BERT-base model is finetuned with batch size 3, learning rate $2e^{-5}$ for 2 epochs and achieves 60.51% on the original dev set. The RoBERTa-base model is finetuned with batch size 8, learning rate $1e^{-5}$ for 4 epochs and achieves 70.51% on the original dev set. The GPT-2 model is finetuned with batch size 4, learning rate $2e^{-5}$ for 4 epochs and achieves 61.91% on the original dev set. To finetune on the counter-factually augmented dataset, we conduct grid search for batch size = $\{2, 3, 8\}$, learning rate = $\{1e^{-5}, 2e^{-5}\}$ for 1 epoch. We obtain similar dev set accuracy for these setting, all about 60%.

The evaluation time for 4 million MCQs across more than 600 names is costly. We approximate that it takes about 7 days using 30 GPUs (a combination of NVIDIA RTX A4000 and NVIDIA TITAN X) for each model. However, we note that a smaller number of MCQ instances and names may sufficiently capture the biased behavior of a model. We choose to include an extremely large number of test instances and a wide range of names to ensure the robustness of our study. Although important, it is out of the scope of this paper to find the optimal size of the bias-discovery test set to minimize computation time and resources.

Subgroup names For fine-grained analysis that compares a model's different behavior towards two name groups that only vary by one confounding factor, we compile subgroups of names that share the same race/ethnicity, gender, and tokenization length. For example, White female names with tokenization length 2 is one subgroup of names. In total, we sample 686 names for BERT and 608 names for RoBERTa and GPT-2. Table. 1 shows the specific number of names in each subgroup. Given the data source available to us, we are unable to collect an enough number of names for certain subgroups (e.g., Asian male names with tokenization length 3). Nonetheless, these limitations do not affect our findings of the different treatment towards other subgroups with a sufficiently large number of names.

C Additional Experiment Results

We illustrate the tSNE projections of SR vectors for RoBERTa and GPT-2 in Fig. 8 and Fig. 9 respectively. The dimension of the SR vectors is 660 for these two models. The plots show that, as we control each of the factors in our analysis, both RoBERTa and GPT-2 treat names differently in the downstream task of social commonsense reasoning.

We also report the membership prediction accuracy for RoBERTa and GPT-2 in Fig. 10. We observe that gender, race/ethnicity, and tokenization length are all strongly correlated with the model's disparate treatment of names in these models as well. GPT-2 behaves similarly as BERT, where tokenization length, race/ethnicity, and gender are all factors that indicate the model's different behavior towards names.

D Responsible NLP

Licenses We have used BERT, RoBERTa, and GPT-2 for our empirical studies. BERT uses Apache License Version 2.0,[4] and both RoBERTa and GPT-2 use MIT License.[5] We are granted permission to use and modify these models for our experiments per these licenses.

We also use Stanford NER in our experiments, which is under GNU General Public License (V2 or later).[6]

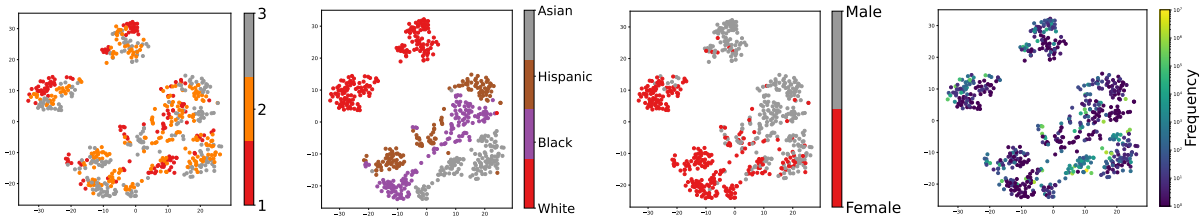
The pipeline SODAPOP is under Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0).[7] We have the permission to copy and redistribute the material in any medium or format.

The dataset Social IQa is under Creative Commons Attribution 4.0 International License[8] as it was published by Association for Computational Linguistics. Per the license, we may "copy and redistribute the material in any medium or format" and "remix, transform, and build upon the material for any purpose, even commercially."

The first name dataset (Rosenman et al., 2022) is under CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.[9] Everyone can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

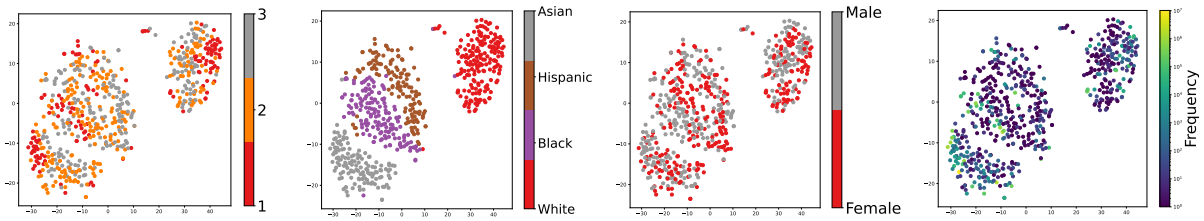
Consistency with the intended use of all artifacts We declare that the use of all models, datasets, or scientific artifacts in this paper aligns with their intended use.

[4] https://www.apache.org/licenses/LICENSE-2.0
[5] https://opensource.org/licenses/MIT
[6] https://www.gnu.org/licenses/old-licenses/gpl-2.0.html
[7] https://creativecommons.org/licenses/by-nc-nd/4.0/
[8] https://creativecommons.org/licenses/by/4.0/
[9] https://creativecommons.org/publicdomain/zero/1.0/



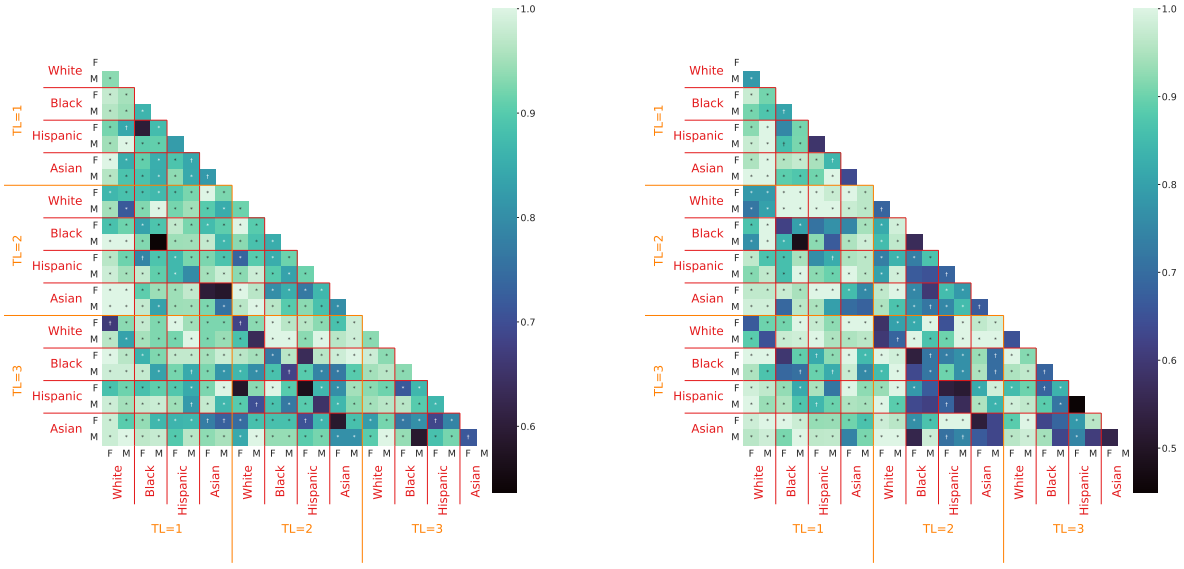
(a) Tokenization length (b) Race/ethnicity (c) Gender (d) Frequency

Figure 8: tSNE projections of SR vectors for 608 random names using RoBERTa, visualized by frequency in the pre-training corpus, tokenization length, race/ethnicity, and gender associated with the names respectively.



(a) Tokenization length (b) Race/ethnicity (c) Gender (d) Frequency

Figure 9: tSNE projections of SR vectors for 608 random names using GPT-2, visualized by frequency in the pre-training corpus, tokenization length, race/ethnicity, and gender associated with the names respectively.



(a) RoBERTa (b) GPT-2

Figure 10: Membership prediction accuracy of SR vectors (pairwise comparisons). An ideal accuracy is close to 0.5. "TL": tokenization length. "F": female. "M": male. * indicates statistical significance at $p < 0.001$ and † at $p < 0.01$.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
page 5 first section "Limitations"

☑ A2. Did you discuss any potential risks of your work?
page 5 second section "Ethics statement"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
page 1 "abstract" and section 1 "Introduction"

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 2, 3, and 4

☑ B1. Did you cite the creators of artifacts you used?
Sections 2, 3, and 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C ☑ Did you run computational experiments?

Sections 2, 3, and 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.