Measuring Social Biases in State Space Models

Mingkai Deng
Carnegie Mellon University
mingkaid@cs.cmu.edu

Jiseung Hong
Carnegie Mellon University
jiseungh@andrew.cmu.edu

To Eun Kim
Carnegie Mellon University
toeunk@cs.cmu.edu

Luca Sfragara
Carnegie Mellon University
lsfragar@cmu.edu

Abstract

Pretrained large language models (LLMs) have demonstrated remarkable capabilities but often exhibit social biases that can reinforce harmful stereotypes. While extensive research has characterized and mitigated biases in Transformer-based LLMs, little is known about the bias properties of emerging architectures such as State Space Models (SSMs). In this work, we systematically evaluate and compare the social biases of SSM-based LLMs, specifically Mamba and Mamba-2, against Transformer-based models with controlled experimental conditions. Using StereoSet and BBQ datasets, we find that Mamba-based LLM exhibits fewer biases but at the cost of weaker language modeling, whereas Mamba-2 offers a more favorable trade-off between bias mitigation and performance. Furthermore, we investigate the impact of debiasing via fine-tuning on the PANDA dataset. Transformer-based models show improved stereotype parity after fine-tuning, albeit with degraded language modeling ability, while Mamba models exhibit limited responsiveness to such debiasing strategies. Our findings suggest that SSM-based models, despite their efficiency advantages, may require fundamentally different approaches—such as counterfactual training or reinforcement learning with human feedback—to effectively mitigate social biases.

1 Introduction

Motivation Foundation models (Bommasani et al., 2021) based on the Transformer (Vaswani, 2017) architecture, such as large language models (LLMs, Radford et al., 2019; Brown et al., 2020), have largely taken over the field of natural language processing (NLP) due to their strong performance and versatility. However, pretrained LLMs typically exhibit significant social biases such as stereotypes (Sheng et al., 2019), which can cause disproportionate harm to marginalized groups. While there is ample research (Gallegos et al., 2024) and robust evaluation frameworks (Ranaldi et al., 2024) for assessing biases in LLMs, they have almost exclusively focused on those with the Transformer architecture, to the detriment of alternatives such as State Space Models (SSMs, Gu et al., 2021a,b).

SSMs, as exemplified by Hyena (Poli et al., 2023) and Mamba (Gu and Dao, 2024), have recently been proposed as alternatives to Transformers for language modeling. These architectures are highly promising due to combining efficient processing of long sequences with the scalability of Transformer-based models. However, little is known about the social biases they may exhibit. One may expect the biases to be worse due to the model being forced to selectively memorize previous content, which may cause it to rely on abstract stereotypes for next-word prediction. Given their potential and the likelihood that such models or hybrid versions (Jelassi et al., 2024) may become widespread, it is crucial to extend bias evaluation and debiasing frameworks to SSMs.

To fill this gap, we study social biases in popular State Space Models (SSMs) and compared them directly with Transformer-based models. Specifically, we focus on Mamba (Gu and Dao, 2024) and Mamba-2 (Dao and Gu, 2024), two representative SSM-based language models. Leveraging publicly available model checkpoints pretrained on identical data and training steps, we rigorously assess their bias characteristics relative to Transformers using the StereoSet (Nadeem et al., 2021) and BBQ dataset (Parrish et al., 2022). Additionally, we explore effective debiasing strategies by fine-tuning these models on the PANDA (Qian et al., 2022) dataset, further deepening our understanding of the potential for reducing social stereotypes across different architectures.

Formally, in this paper, we address the following research questions (RQs):

- **RQ1**: Could the architectural difference between transformer-based and SSM-based model lead to different bias trend?

- **RQ2**: Similar to the transformer-based models, do SSM-based models show a tradeoff between language modeling performance and stereotypical bias issues?

- **RQ3**: Compared to transformer-based models, do SSMs exhibit greater reductions in stereotypical bias when finetuned on demographically diverse data?

Thesis Statement This project investigates whether the architectural differences of State-Space Models (SSMs), which selectively memorize past tokens, lead to distinct bias patterns compared to Transformer-based models, examines if the correlation between improved language modeling performance and higher stereotype bias observed in Transformers also applies to SSMs, and evaluates whether SSMs exhibit greater reductions in stereotypical biases when fine-tuned on demographically diverse data.

Bias Statement Transformer-based pre-trained language models (LLMs) often exhibit ethical concerns, particularly social biases related to gender, race, profession, and religion, which can reinforce harmful stereotypes and disproportionately impact marginalized groups by influencing AI-driven decisions in unfair ways. An ideal unbiased model would ensure that socially sensitive attributes do not systematically affect predictions, promoting fair and equitable language generation across all demographics. In this work, we compare the architectural biases of State Space Models (SSMs) and Transformer-based models, recognizing that SSMs may amplify stereotypes due to their selective memory mechanisms; ideally, a model should assign equal likelihood to stereotypical and anti-stereotypical outputs.

2 Related Work

Social Biases in Foundation NLP Models Foundation language models (Bommasani et al., 2021) have revolutionized how NLP tasks are performed. Despite their impressive text processing power, they are known to exhibit significant social biases (Sheng et al., 2019; Gallegos et al., 2024) and run the risk of causing disproportionate harm. While recent works study and attempt to mitigate the biases of pretrained LLMs (Gallegos et al., 2024; Ranaldi et al., 2024), they have primarily focused on Transformer-based models, leaving emerging architectures understudied.

State Space Models State Space Model (SSM) is an emerging class of neural network architecture that combine the scalability of Transformer and efficiency of Recurrent Neural Networks for modeling sequential data (Poli et al., 2023; Gu and Dao, 2024). Recent work such as Mamba and Mamba-2 (Gu and Dao, 2024; Dao and Gu, 2024) demonstrate favorable scaling laws relative to typical Transformer solutions, but the evaluations have typically focused on loss metrics such as perplexity and task-specific scores such as QA accuracy, without adequate accounting for model biases.

3 Methods

3.1 Evaluation of Pretrained Checkpoints

Task Definition As the first task, we aim to compare existing stereotypical/social bias inherent in pretrained transformer-based models and SSM-based models to answer the first two research questions (RQ1 and RQ2). Therefore, we keep the other conditions other than the architectural difference—such as pretrained data and training methods—the same. This ensures we can attribute the evaluation results to the architectural difference. As evaluation datasets, we use StereoSet (Nadeem et al., 2021) and Bias Benchmark for QA (BBQ) (Parrish et al., 2022) on both transfomer-based LM and SSM-based LM and adopt their metrics introduced in the paper.

3.1.1 Evaluation with StereoSet

StereoSet is (Nadeem et al., 2021) for measuring stereotypical biases in four domains: gender, profession, race, and religion. For metrics, it consists of two types of evaluations: intra-sentence and inter-sentence tests. Following Ranaldi et al. (2024), we adopt intra-sentence tests for our initial experiments. Inter-sentence tests require model fine-tuning, introducing an additional potential source of bias that is difficult to control.

After inference completion, we store the predicted sentence completions generated by the model and evaluate them using established bias metrics. Specifically, we adopt the following four metrics:

Stereotype Score (SS) The Stereotype Score (SS) measures how frequently a model prefers a stereotypical sentence over an anti-stereotypical one. Below is the definition of the raw SS:

$$\text{SS-Raw} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(P_{\text{stereo},i} > P_{\text{anti-stereo},i}\right) \quad (1)$$

where:

- $P_{\text{stereo},i}$ is the probability assigned to the stereotypical sentence,
- $P_{\text{anti-stereo},i}$ is the probability assigned to the anti-stereotypical sentence,
- $\mathbb{I}(\cdot)$ is the indicator function (1 if true, 0 otherwise),
- N is the total number of stereotype-anti-stereotype pairs.

As the ideal LLM should have balanced preference between the stereotypical and non-stereotypical sentences, we transform SS-Raw into the balanced SS as below:

$$\text{SS} = \min(\text{SS-Raw}, 100 - \text{SS-Raw}) \times 2 \quad (2)$$

Language Modeling Score (LMS) The Language Modeling Score (LMS) assesses how often the model prefers meaningful sentences over nonsensical ones:

$$\text{LMS} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(P_{\text{stereo/anti-stereo},i} > P_{\text{nonsense},i}\right) \quad (3)$$

where:

- $P_{\text{stereo/anti-stereo},i}$ is the probability assigned to either the stereotypical or anti-stereotypical sentence,
- $P_{\text{nonsense},i}$ is the probability assigned to the nonsensical sentence.

We then normalize the SS score so that it ranges from 0 to 100, with 100 being the best, that is, the least biased. This means that a perfect model with a SS score of 100 would always randomly pick between the stereotypical and anti-stereotypical continuation.

Idealized CAT Score (ICAT) The Idealized CAT (ICAT) Score balances stereotype bias and language modeling performance:

$$\text{ICAT} = \text{LMS} \times \text{SS} \times \frac{1}{100} \quad (4)$$

Overall Score The overall score for a model is computed as:

$$\text{Overall Score} = \text{ICAT} \times 100 \quad (5)$$

This transformation converts the ICAT score into a percentage scale for easier interpretation.

3.1.2 Evaluation with BBQ

The Bias Benchmark for QA (BBQ) dataset (Parrish et al., 2022) assesses social biases in question-answering (QA) models by providing contextually implied demographic cues within contrasting scenarios. Each scenario consists of pairs of contexts, one stereotypical and one anti-stereotypical, designed to test whether models can avoid bias based on demographic attributes such as race, gender, or socioeconomic status. An illustrative example involves comparing model responses when presented with two similar situations featuring individuals from different demographic groups, ensuring that the model does not preferentially select answers driven by stereotypical associations.

Accuracy The primary evaluation metric used in the BBQ dataset is accuracy, which measures whether a model selects the correct answer in scenarios explicitly constructed to highlight potential biases. This metric emphasizes performance on contrastive examples, where biases might significantly influence the model's choice. Accurate performance indicates the model's capability to respond fairly and neutrally, without succumbing to implicit social biases embedded within stereotypical contexts.

3.2 Model Finetuning for Debiasing LMs

Task Definition As the second task, and to answer the RQ3, we finetune both transformer-based and SSM-based LMs with demographically diverse data. To form such a dataset, we augment the PANDA dataset (Qian et al., 2022).

3.2.1 PANDA

The PANDA dataset (Qian et al., 2022) consists of approximately 98,000 sentence pairs, each including an original sentence and its correspond-

ing human-annotated rewrite designed to alter demographic references. These perturbed sentences specifically target demographic variations in attributes such as race and gender, enabling the evaluation and mitigation of biases in language models. We augment their dataset by diversifying sentence structures for training and finetune an LM by causal language modeling task.

3.3 Training Example Diversification for Bias Mitigation

Direct diversity exposure We simply construct an instance only with a rewritten sentence.

Multiple demographic pairs comparison Whenever possible, we construct up to three rewritten variants of an original sentence, explicitly showing demographic variations. The sentence is structured as follows: *Text: {original}. More inclusive version of the text: {rewrite}.*

Demographic knowledge exposure Rather than explicitly showing rewritten sentences, we construct a data instance with pedagogical sentence about demographics as follows: *The statement {original} includes a reference to {word}. This can be represented in different ways to be inclusive of various demographics, including {category} perspectives.*

Demographic examples without instructions This time, we only show rewritten sentences, and do not reveal an original sentence, aiming for implicit learning by causal language modeling. The sentence is structed as follows: *Here is a statement that uses inclusive language for {category} demographics: {rewrite}.*

4 Experimental Setup

4.1 Evaluation of Pretrained Checkpoints

Because LLM biases and language-modeling performance are not only dependent on architecture, but also training data, setup, and model size, we control for these confounding variables strictly by evaluating pretrained checkpoints that are exactly the same across all aspects except for architecture. Specifically, we evaluate `state-spaces/trasformerpp-2.7b`, `state-spaces/mamba-2.8b`, and `state-spaces/mamba2-2.7b`, which are released by the Mamba authors (Gu and Dao, 2024), are of comparable sizes of 2.7-2.8B parameters, and are all trained on the Pile (Gao et al., 2020) dataset (300B tokens) for the same number of steps. We evaluate these models by measuring the probability of sentences from StereoSet using the Mamba codebase [1] and computing the metrics defined in Sec. 3.1.

4.2 Model Finetuning

For the same reason as in Sec. 4.1, we finetune models with identical training data, setup, and model size. Because the original Mamba codebase does not support finetuning, we use Huggingface Transformers for this part of the experiments and finetune model checkpoints compatible with it, namely `EleutherAI/pythia-2.8b` and `state-spaces/mamba-2.8b-hf`, which are also trained on the same Pile dataset with exactly the same setting. We finetune the two models on the PANDA dataset for 1 epoch with an effective batch size of 64, learning rate of 10^{-6}, and `bfloat16` precision. The finetuning uses 2 NVIDIA L40S GPUs and takes 3 hours for `pythia-2.8b` and 4.5 hours for `mamba-2.8b-hf`. We evaluate the finetuned models using the same procedure as in Sec. 4.1.

5 Results and Analysis

5.1 Evaluation of Pretrained Checkpoints

RQ1*: Could the architectural difference between transformer-based and SSM-based model lead to different bias trend?*

Figures 1 and 2 show the results obtained from evaluating TransformerPP, Mamba and Mamba 2 Model on StereoSet and BBQ (full numerical results presented in Tables 1 and Table 2 in Appendix §A. Regarding Stereoset, the first interesting analysis is that while all three models exhibit a similar language score (they can model language equally well), albeit with Mamba 2 being slighly better, we observe a wider variation in the SS score. Mamba is the least biased model, followed by Mamba 2 and Transformer. This results also holds across all the 4 categories (gender, race, profession and religion).

RQ2*: Similar to the transformer-based models, do SSM-based models show a tradeoff between language modeling performance and stereotypical bias issues?*

We performed further analysis on the correlation between the LM and SS score across the different models and categories, as we were interested in the

[1] `www.github.com/state-spaces/mamba`

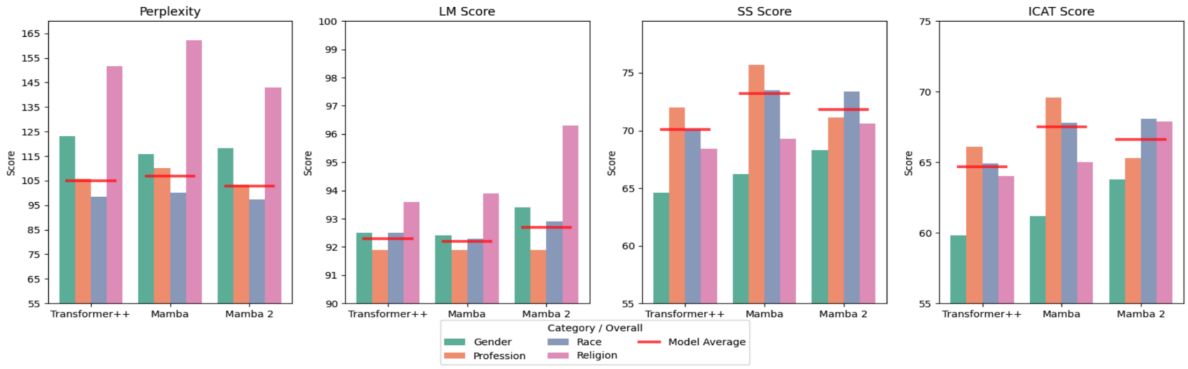

Figure 1: Bar charts for StereoSet Results for Transformer, Mamba and Mamba-2 Models

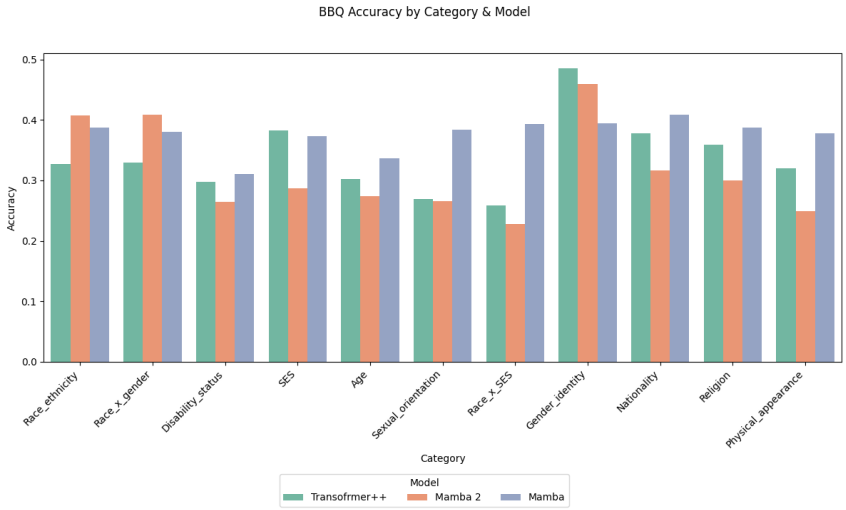

Figure 2: Bar chart for BBQ evaluation accuracy for Transformer, Mamba and Mamba-2 by category

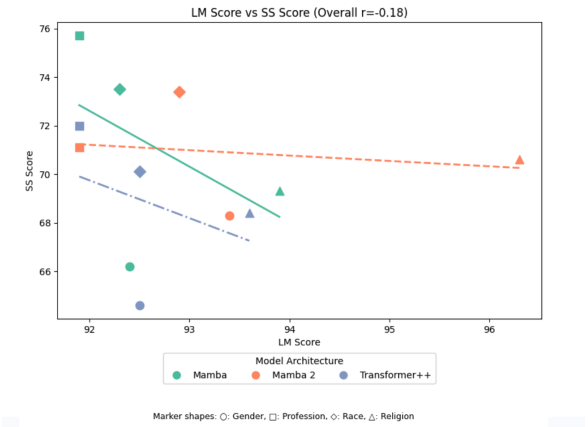

Figure 3: Correlation Results between Language Modelling and Stereotype Score

known tradeoff (Omrani Sabbaghi et al., 2023) between an LLM's ability to model language and bias. Sabbaghi et al. show that as you give the model more parameters and data to reduce its language-modeling loss, it becomes better at capturing every statistical regularity in the data—including the unwanted ones. Our hypothesis aligns with the literature, suggesting that better LM ability comes at the expense of higher exposure to bias, at least in non-finetuned and smaller models as those we are analyzing. This is confirmed by our limited sample, which exhibits a mean negative correlation of 0.18 between the LM and SS score across the 3 models. Interestingly, this correlation was about -0.10 for Mamba Models, indicating a promising lower tradeoff between bias and ability to model language. These results are visualized in Figure 3 To increase the robustness of our analysis, we extend it to the BBQ dataset, which is more comprehensive (11 categories against only 4 of Stereoset). While we can only test the equivalent of the SS score, we find similar results. Mamba models are overall less biased than transformer, achieving a higher accuracy in 9 out of 11 categories.

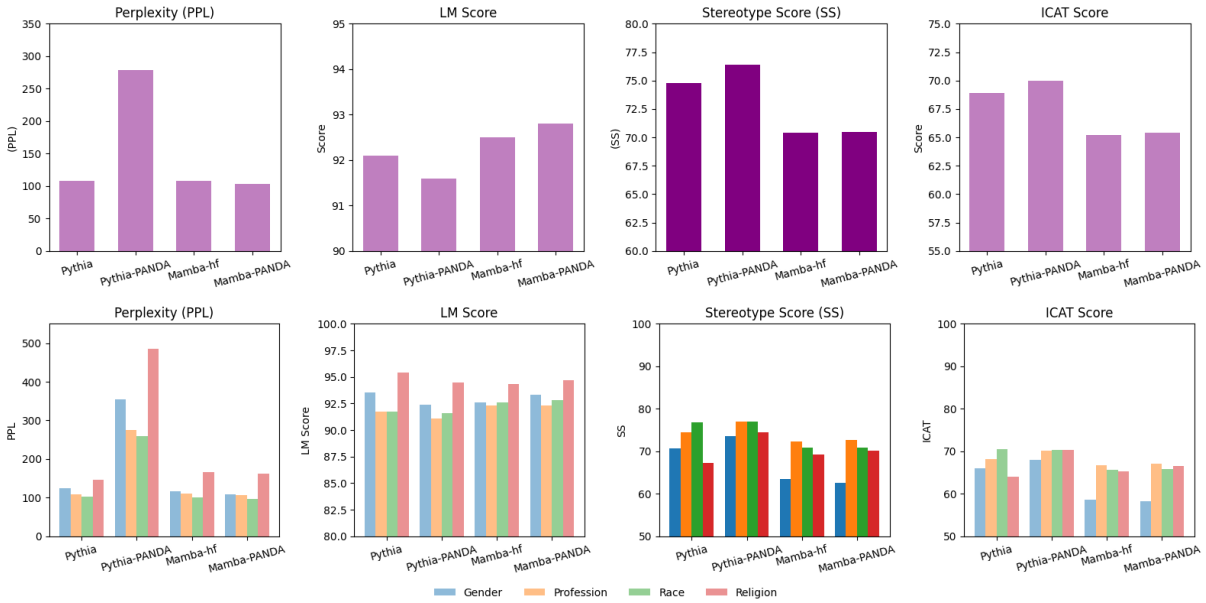

Figure 4: Bar charts for StereoSet results overall (top) and segmented by domain (bottom) for Pythia, Pythia-PANDA, Mamba-hf, and Mamba-PANDA models, including perplexity (ppl), stereotype score (ss), internal consistency (icat), and language modeling score (lms).

5.2 Model Finetuning for Debiasing LMs

RQ3: Compared to transformer-based models, do SSMs exhibit greater reductions in stereotypical bias when finetuned on demographically diverse data?

Figure 4 shows the results from finetuning the model on the PANDA dataset (full numerical results presented in Table 3 in Appendix §B). Transformer-based Pythia trades off language modeling ability for improved stereotype parity, as witnessed by the higher Stereotype score (74.8 → 76.4) at the cost of higher Perplexity (108.5 → 278.7). On the other hand, Mamba model shows little performance change given the same training setting and data, with only marginally higher SS (70.4 → 70.5) but lower PPL (108.1 → 103.9). Segmented analysis by stereotype categories further reveals interesting patterns in mitigation transfer for the two architectures. While PANDA covers only the Gender and Racial stereotypes, for the Transformer-based Pythia, the SS improvement mitigation applies even for types not covered by PANDA. For instance, aside from Gender with 70.6 → 73.6 and Race with 76.8 → 76.9, Profession and Religion also improve with 74.4 → 77.0 and 67.2 → 74.5, respectively. On the other hand, Mamba is strangely not as responsive to finetuning as a debiasing approach, as the SS becomes even worse for Gender (63.5 → 62.5) and barely changes for Race (70.9 → 70.9). However, SS on Profession and Religion improve slightly with 72.3 → 72.7 and 69.3 → 70.2, respectively.

6 Discussion and Insights

As hypothesized at the outset of our project, SSM-based LLMs (e.g., Mamba and Mamba-2) show different patterns of stereotypes and responses to debiasing strategies. Preliminary results show that Mamba has fewer stereotypes at the cost of weaker language modeling, while Mamba-2 shows better trade-off between LM performance and stereotypes. Furthermore, finetuning experiments on augmented data like PANDA show that the Mamba-based LLM is comparatively unresponsive to finetuning, while the Transformer-based LLM shows fewer stereotypes and transfers across categories after finetuning at the cost of worse language modeling performance, as often observed with models with similar architectures (Omrani Sabbaghi et al., 2023). This may be due to Mamba having learned patterns for selective memory (Gu and Dao, 2024) from pretraining that is difficult to change with short finetuning. Despite efficiency advantages, SSM-based LLMs may require different strategies to debias properly, such as more decoding-time methods.

7 Limitations

Our study has limited demographic coverage, focusing exclusively on the PANDA dataset categories of gender and race, and explores only finetuning with augmented datasets for debiasing, neglecting methods such as decoding-time approaches. Additionally, the passive learning approach through causal language modeling may inadequately capture nuanced stereotype reasoning, highlighting a need for active learning strategies with targeted feedback.

Potential Risks One potential risk in this work lies in the limitations of existing bias metrics, which often reduce complex and context-dependent social biases to simplified scalar scores. This reductionism may obscure subtle or intersectional forms of bias, leading to incomplete or misleading conclusions. Additionally, evaluation datasets such as StereoSet may themselves encode cultural or annotator-induced biases, which can confound the attribution of bias to the model architecture. As a result, observed disparities may reflect dataset artifacts rather than intrinsic model behavior, underscoring the importance of careful interpretation and dataset validation in bias assessment.

8 Conclusion and Future Work

Conclusion We perform one of the first systematic investigations into social biases in State Space Model (SSM) based language models compared to traditional Transformer architectures. We find that while SSMs like Mamba and Mamba-2 show promising bias properties and efficiency advantages, they exhibit different trade-offs between language modeling ability and stereotype expression. Furthermore, standard finetuning methods effective for Transformers appear less effective for SSMs, suggesting that architectural differences fundamentally shape how models internalize and express social biases. These results highlight the need for architecture-specific evaluation and mitigation strategies as the diversity of foundation model designs continues to grow.

Future Work Future work involves conducting a broader range of bias evaluations using generation-based methods instead of solely relying on likelihood-based metrics. Generation-based evaluations better reflect real-world model usage scenarios, as biases often manifest prominently during active text generation. Evaluating models based on their generated outputs allows for a more comprehensive understanding of how biases are expressed in practical applications.

Additionally, exploring counterfactual reasoning is crucial for improving demographic representation in State-Space Models (SSMs). Current methods primarily focus on passive exposure to stereotype-related data, which might limit the models' ability to grasp subtle differences and actively challenge stereotypes. By training models explicitly on counterfactual scenarios—such as how altering demographic attributes changes outcomes—SSMs can develop more nuanced reasoning capabilities, ultimately reducing their susceptibility to reinforcing harmful stereotypes.

Finally, applying reinforcement learning with human feedback (RLHF), guided by targeted evaluation metrics like the stereotype score (SS), presents a promising direction for enhancing debiasing strategies in SSM-based models. Current finetuning approaches might not effectively correct deeply ingrained biases, as evidenced by limited responsiveness in some model architectures. Integrating RLHF enables active learning from human judgments or metric-based feedback, encouraging models to internalize fairness criteria explicitly and to adapt their behavior dynamically, ultimately leading to more robust and socially responsible language models.

References

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

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

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

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

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Albert Gu, Karan Goel, and Christopher Ré. 2021a. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. 2021b. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. 2024. Repeat After Me: Transformers are Better than State Space Models at Copying. *arXiv preprint arXiv:2402.01032*.

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. Association for Computational Linguistics.

Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, pages 1–12, Montréal, QC, Canada.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *Preprint*, arXiv:2205.12586.

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.

Leonardo Ranaldi, Elena Sofia Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Massimo Zanzotto. 2024. A Trip Towards Fairness: Bias and De-Biasing in Large Language Models. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 372–384, Mexico City, Mexico. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

A Evaluation of pretrained checkpoints

Table 1 and 2 show pretrained checkpoints evaluation results on StereoSet and BBQ.

B Model Finetuning for Debiasing LMs

Table 3 shows finetuned/debiased models evaluation results on StereoSet.

C Computational Costs

As mentioned in §4.2, the finetuning process is done with 2 NVIDIA L40S GPUs which took 3 hours for `pythia-2.8b` and 4.5 hours for `mamba-2.8b-hf`.

Domain	Model	*ppl*	*ss*	*icat*	*lms*
Overall	state-spaces/transformerpp-2.7b	104.9	70.1	64.7	92.3
	state-spaces/mamba-2.8b	**108.1**	70.4	65.2	92.5
	state-spaces/mamba2-2.7b	102.8	**71.8**	**66.6**	**92.7**
Gender	state-spaces/transformerpp-2.7b	123.1	64.6	59.8	92.5
	state-spaces/mamba-2.8b	117.5	63.4	58.7	92.6
	state-spaces/mamba2-2.7b	118.2	68.3	63.8	93.4
Profession	state-spaces/transformerpp-2.7b	105.7	72.0	66.1	91.9
	state-spaces/mamba-2.8b	111.5	72.3	66.7	92.3
	state-spaces/mamba2-2.7b	103.3	71.1	65.3	91.9
Race	state-spaces/transformerpp-2.7b	98.5	70.1	64.0	92.5
	state-spaces/mamba-2.8b	100.9	70.9	65.6	92.6
	state-spaces/mamba2-2.7b	97.4	73.4	68.1	92.9
Religion	state-spaces/transformerpp-2.7b	151.7	68.4	64.0	93.6
	state-spaces/mamba-2.8b	166.5	69.3	63.3	94.3
	state-spaces/mamba2-2.7b	143.0	70.6	67.9	96.3

Table 1: StereoSet results across domains for different models, including perplexity (ppl), stereotype score (ss), internal consistency (icat), and language modeling score (lms).

Category	transformerpp-2.7b	mamba-2.8b	mamba2-2.7b
Overall	0.337	0.353	**0.381**
Race	0.328	**0.407**	0.387
Race+Gender	0.330	**0.408**	0.380
Disability	0.298	0.264	**0.311**
SES	**0.383**	0.287	0.373
Age	0.302	0.274	**0.336**
Sexual Orientation	0.270	0.266	**0.384**
Race+SES	0.258	0.228	**0.394**
Gender Identity	**0.486**	0.459	0.395
Nationality	0.378	0.317	**0.409**
Religion	0.359	0.300	**0.388**
Physical Appearance	0.320	0.249	**0.378**

Table 2: BBQ evaluation accuracy by category.

Domain	**Model**	*ppl*	*ss*	*icat*	*lms*
Overall	pythia	108.5	74.8	68.9	92.1
	pythia-panda-full-ft	278.7	76.4	70.0	91.6
	mamba-hf	108.1	70.4	65.2	92.5
	mamba-panda-full-ft	103.9	70.5	65.4	92.8
Gender	pythia	125.3	70.6	66.0	93.5
	pythia-panda-full-ft	355.3	73.6	68.0	92.4
	mamba-hf	117.5	63.4	58.7	92.6
	mamba-panda-full-ft	108.9	62.5	58.3	93.3
Profession	pythia	109.4	74.4	68.2	91.7
	pythia-panda-full-ft	274.6	77.0	70.1	91.1
	mamba-hf	111.5	72.3	66.7	92.3
	mamba-panda-full-ft	107.3	72.7	67.1	92.3
Race	pythia	102.7	76.8	70.5	91.7
	pythia-panda-full-ft	258.3	76.9	70.4	91.6
	mamba-hf	100.9	70.9	65.6	92.6
	mamba-panda-full-ft	97.6	70.9	65.8	92.8
Religion	pythia	145.5	67.2	64.1	95.4
	pythia-panda-full-ft	484.7	74.5	70.4	94.5
	mamba-hf	166.5	69.3	63.3	94.3
	mamba-panda-full-ft	163.1	70.2	66.5	94.7

Table 3: StereoSet results across domains for Pythia, Pythia-PANDA, Mamba-hf, and Mamba-PANDA models, including perplexity (ppl), stereotype score (ss), internal consistency (icat), and language modeling score (lms).