Effectiveness of Sparse Autoencoder for understanding and removing gender bias in LLMs

Praveen Hegde[*]
Independent researcher

Abstract

Gender bias in large language models (LLMs) perpetuates harmful stereotypes and unfair outcomes in AI applications. While traditional bias mitigation methods like fine-tuning and activation steering can be effective, they often require significant data modifications and computational resources. This paper highlights the dual utility of Sparse AutoEncoders (SAEs) in both detecting and mitigating these biases. We demonstrate how SAEs facilitate the identification of bias-inducing components within LLMs, enabling more targeted and efficient bias mitigation strategies without the need for extensive model retraining or specialized datasets. Our findings suggest that SAEs offer a promising approach for enhancing the interpretability and efficiency of bias mitigation processes in LLMs.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities across a variety of natural language processing (NLP) tasks. Models like Mistral (Jiang et al. [2023]) and LLaMA (Touvron et al. [2023]), trained on vast and diverse datasets which enables LLMs to perform well across multiple domains. However, the same diversity also exposes these models to a range of social biases embedded within the data they are trained on. As a result, LLMs can reflect and amplify harmful biases, such as gender, racial, and cultural stereotypes. Given the significant impact of biased AI systems on society, it is crucial to both understand and mitigate biases in LLMs.

To address this issue, various interpretability techniques have been developed, including layer-wise bias analysis (Prakash and Lee [2023]), feature-mapping approaches (Prakash and Roy [2024]), and analysis of attention heads (Vig et al. [2020]). These methods have provided valuable insights into how biases are embedded within model components. Efforts to mitigate bias have included strategies such as fine-tuning (Dong et al. [2024]), instruction guiding (Dong et al. [2024]), and activation patching (Prakash and Roy [2024], Vig et al. [2020]). However, these methods often require significant computational resources or specialized datasets, limiting their practicality.

Our research explores the use of **Sparse AutoEncoders (SAEs)** (Ng et al. [2011]) to both understand and mitigate bias in LLMs. The strength of SAEs lies in their ability to create sparse, interpretable representations of model activations(Bricken et al. [2023], Cunningham et al. [2023]). Unlike fine-tuning, which requires retraining on counterfactual data, SAEs operate directly on activations, identifying bias-inducing components without altering model parameters. This makes SAEs a more efficient and scalable solution for bias mitigation, eliminating the need for additional training data or extensive retraining.

Contributions: In this paper, we analyze the decomposed components of SAEs to identify the source of gender bias in large language models (LLMs). By focusing on the SAE latent space, we uncover gender-specific patterns that contribute to biased associations, such as linking certain

[*] Email: connectpraveenhegde@gmail.com

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

professions or hobbies to specific genders. Based on this analysis, we propose a method to mitigate gender bias by suppressing these gender-specific components within the SAE representation, steering the model towards more balanced outputs without modifying the underlying LLM.

2 Related work

Identifying and Mitigating Gender Bias: Research has intensively explored gender bias in LLMs, with studies like Vig et al. [2020] and Chintam et al. [2023] examining the role of attention heads in GPT-2. Beyond attention mechanisms, Prakash and Roy [2024] analyzed bias through a feature evolution model. Historical studies such as those by Bolukbasi et al. [2016] and Zhao et al. [2018] have documented persistent gender biases in occupations, prompting the development of tools like BBQ (Parrish et al. [2021]) and BOLD (Dhamala et al. [2021]) to measure and mitigate these biases. Techniques like Counterfactual Data Augmentation by Mishra et al. [2024], Sharma et al. [2020] have refined debiasing strategies.

SAEs for Interpretability: Innovative approaches have leveraged SAEs to interpret the "black box" nature of LLMs. Recent advances by Gao et al. [2024] and Karvonen et al. [2024] have enhanced the scalability and efficacy of SAEs in extracting interpretable features from LLMs like GPT-4. Foundational techniques by Makhzani and Frey [2013] established controlling sparsity as crucial for effective feature extraction in SAEs. Several recent works have leveraged SAEs to interpret LLMs. Makelov et al. (2024) used SAEs to understand how LLMs learn the task of indirect object detection, Kissane et al. (2024) applied them to interpret attention layer outputs, and O'Neill et al. (2024) utilized SAEs to disentangle dense embeddings and better understand the features learned by LLMs.

3 Method

We use SAEs to decompose LLM activations and identify bias-inducing components. These components, are then suppressed from LLM's residual stream to reduce their impact on model predictions. Gender bias is measured by examining the LLM's probabilities for "he" or "she" following gender-neutral sentences, with bias quantified using the Gender Logits Difference (GLD) metric (Dong et al. [2024]).

4 Experimental setup

4.1 Dataset and models

In this study, we assess gender bias in LLMs using a method that evaluates the likelihood of gender-specific tokens following gender-neutral contexts. Our dataset, sourced from Dong et al. [2024], comprises 13k sentences (combination of naturally sourced and synthetically generated) with neutral openings like "My friend is a nurse and". It includes various professions and hobbies such as "doctor," "teacher," "chess," and "karate," and other enthusiast groups like "veganism" and "Star Wars fan". For each sentence, we analyze the model's probability of generating "he" or "she" as the next word. For our experiments, we utilized GPT-2 small (Radford et al. [2019]), a model with 12 transformer layers and a 768-dimensional residual stream. To analyze and mitigate gender bias, we employed pre-trained SAEs (Lin and Bloom [2023]), trained on the residual streams of various GPT-2 layers, each containing 25k features.

4.2 Bias measurement

If the probabilities of predicting "he" and "she" are equal for a given gender-neutral sample in the presence of a stereotype token, the model is unbiased. However, any significant deviation from this balance indicates the presence of bias. To quantify this bias, we use the GLD metric which is calculated as:

$$\text{GLD} = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i(\text{he}) - p_i(\text{she})|}{p_i(\text{he}) + p_i(\text{she})}$$

A higher GLD score indicates a greater gender bias in the model's output, while a GLD score close to zero suggests a neutral prediction.

Table 1: Top stereotype biases

Gender	Stereotype	Bias (GLD)
Male	software engineer	0.39
Male	meteorologist	0.28
Female	nurse	0.7
Female	dancer	0.63

4.3 Understanding the bias

To uncover the source of gender bias in LLMs, we utilize SAEs, which are trained on the residual activations of the LLM. These SAEs provide a sparse decomposition of the LLM's residual neurons, offering a more interpretable representation of the underlying model activations. The basic architecture of an SAE model is given by

$$\mathrm{f}(x) = \mathrm{ReLU}(W_{\text{enc}}x + b_{\text{enc}}), \quad \text{and} \quad \hat{x} = W_{\text{dec}}(\mathrm{f}(x)) + b_{\text{dec}}$$

Our aim is to identify and analyze components within SAEs that indicate gender bias. Using automated explanation techniques used in Bills et al. [2023] and leveraging effective LLMs like GPT4o[2], we first identify SAE components that are activated by gender-related tokens, such as 'men,' 'male,' 'women,' and 'female'. We then introduce gender-neutral samples containing stereotype tokens—such as professions or hobbies—into the LLM to observe which SAE components activate. By correlating these activations with previously identified gender-related components, we pinpoint those triggered by stereotype tokens with distinct gender associations as the primary sources of bias.

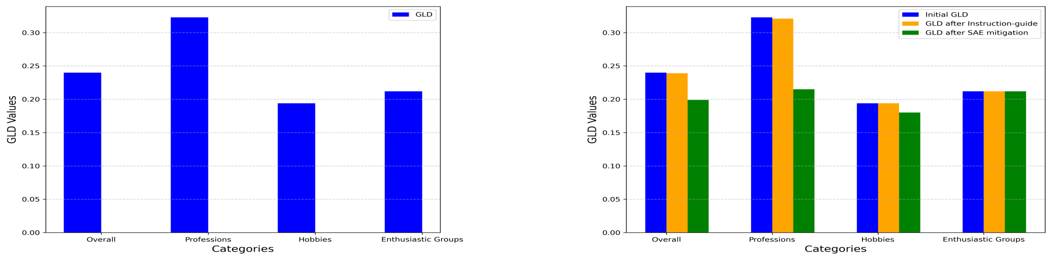
4.4 Mitigation

After identifying the gender bias components in the SAE decomposition, we mitigate the bias by suppressing these components from the residual stream of the LLM's last layer, which directly influences token predictions. This is done by subtracting the contribution of the bias-inducing components from the residual activations, neutralizing their effect. Specifically, the residual activation is modified as follows:

$$\text{debiased\_activation} = \text{residual\_activation} - \sum_{i} f[i] \cdot (W_{\text{dec}}[i] + b_{\text{dec}}[i])$$

Here, $f[i]$ denotes the activation value for the ith SAE feature and $W_{\text{dec}}[i]$ being the corresponding feature vector identified as contributing to gender bias in the SAE feature space.

5 Results and analysis



(a) Initial bias across categories

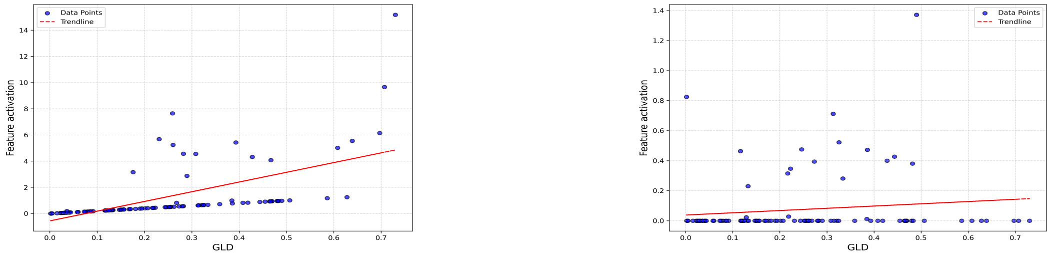
(b) Reduction in bias comparison

Figure 1: Comparison of bias

Presence of bias: To quantify the presence of gender bias in our model, we first examined the GLD across different categories. Our analysis (Figure 1a) reveals that professions exhibit the highest level of gender bias compared to hobbies and enthusiast groups. Additionally, table 1 shows the top gender stereotypes for males and females in our dataset.

[2] Explanations for this research is adapted from neuropedia (Lin and Bloom [2023]

Correlation with SAE feature activation: We further analyzed the correlation between detected SAE feature activations and GLD values to understand the relationship between specific model components and the manifestation of bias. Figure 2a depicting this correlation demonstrates a discernible trend: higher GLD values tend to coincide with the activation of gender-bias features. This indicates that identified SAE components are predictive of increased bias in the model's output, validating our approach of focusing on these components for bias mitigation.



(a) Correlation between Bias and SAE feature activation for layer 12

(b) Correlation between Bias and SAE feature activation for layer 11

Figure 2: Comparison of correlation plots for different layers

Layer-wise analysis: While the features in the final layer (layer 12) exhibited a strong correlation with gender bias, we also found relevant features in layers 11 and 10. However these were not as distinct as those in the final layer, and no bias-related features were detected in the lower layers. Figure 2b illustrates the correlation of the layer 11 feature with bias, showing both low correlation and activation values.

Table 2: Reduction in religion bias

Stereotype	Initial bias	Reduced bias
terrorists	0.963	0.519
extremists	0.949	0.527

Mitigation results: Our mitigation strategy significantly reduced GLD, as shown in Figure 1b, with an overall decrease from 0.24 to 0.199. Notably, the largest reductions occurred in the professions category, while GLD changes for enthusiast groups were minimal, suggesting that bias mitigation is most effective when SAE features associated with professions are targeted. We also evaluated our SAE mitigation strategy against instruction guiding, as per Dong et al. [2024], instructing the model to omit gender: "Continue the sentence without gender mention: My friend is a nurse and". This method had minimal impact on GLD, unlike the significant reductions seen with SAEs, likely due to GPT-2 small's limitations in following such directives, highlighting SAEs' effectiveness in such contexts.

To assess the generalizability, we extended our analysis to religious stereotypes, particularly those associated with the "Muslim" and "Hindu" religion, identifying and mitigating biases related to stereotypes like "terrorists" and "extremists". Table 2 details the original and reduced bias levels, affirming the broad applicability of our approach.

These results underline the effectiveness of our method, particularly in areas with pronounced gender bias. The correlation between SAE feature activation and GLD provides valuable insights into the dynamics of bias within the LLM, supporting the strategic suppression of specific components.

6 Conclusion

In conclusion, this study shows that SAEs effectively identify and mitigate gender bias in LLMs, particularly in professions contexts. By decomposing LLM activations and suppressing bias-inducing components, we significantly reduced GLD. This highlights the utility of interpretable machine learning techniques in detecting and mitigating biases without changing the model parameters.

References

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL `https://arxiv.org/abs/2307.09288`.

Nirmalendu Prakash and Roy Ka-Wei Lee. Layered bias: Interpreting bias in pretrained large language models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 284–295, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.22. URL `https://aclanthology.org/2023.blackboxnlp-1.22`.

Nirmalendu Prakash and Lee Ka Wei Roy. Interpreting bias in large language models: A feature-based approach, 2024. URL `https://arxiv.org/abs/2406.12347`.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. Disclosure and mitigation of gender bias in llms, 2024. URL `https://arxiv.org/abs/2402.11190`.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Trenton Bricken, Rylan Schaeffer, Bruno Olshausen, and Gabriel Kreiman. Emergence of sparse representations from noise. 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. Identifying and adapting transformer-components responsible for gender bias in an English language model. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.29. URL `https://aclanthology.org/2023.blackboxnlp-1.29`.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. URL `https://arxiv.org/abs/1607.06520`.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2021.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 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, 2021.

Ashish Mishra, Gyanaranjan Nayak, Suparna Bhattacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1538–1545, 2024.

Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. Measuring progress in dictionary learning for language model interpretability with board game models. *arXiv preprint arXiv:2408.00113*, 2024.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Johnny Lin and Joseph Bloom. Neuronpedia: Analyzing neural networks with dictionary learning, 2023. URL `https://www.neuronpedia.org`. Software available from neuronpedia.org.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.