USING CONTRASTIVE ACTIVATION ADDITION TO COMBAT SOCIETAL BIASES IN LANGUAGE MODELS

INTRODUCTION

CONTRIBUTION

METHODOLOGY

PURPOSE

ANALYSIS

CONCLUSIONS

REFERENCES

HYPOTHESIS

Today's AI large language models are being used for more important decisions than ever, such as writing legal documents, administering medicine, and hiring job applicants.

However, various studies have shown that they have societal biases about race (Yang et al., 2024) and gender (Kotek et al., 2023). This project uses interpretability-based techniques to reduce the effect of these biases on model responses.

Recently, interpretability-based techniques such as contrastive activation addition (CAA) have shown promise for steering models towards behaviors (Zou et al., 2023).

Previous work has done the following:

- Panickssery et al. (2024) found sycophancy vectors
- Zou et al. (2023) found honesty vectors and neurons corresponding to dishonesty
- Lu & Rimsky (2024) found bias vectors

My work builds on this by identifying bias-correlated neurons. It also develops a new real-world bias benchmark. AI companies and future researchers could use this benchmark for their models, and they can consider CAA as a bias mitigation strategy based on results.

1. Develop a Benchmark

Develop a benchmark with real-world scenarios (using legal, medical, and job hiring cases).

2. Use behavior-positive prompts

Prompt the model with text that matches the desired behavior, e.g. *"All people should get treated fairly and equally regardless of race."* Measure model activations with these prompts at chosen layer.

3. Use behavior-negative prompts

Prompt the model with text that opposes the desired behavior. Measure the model activations with these prompts at chosen layer.

4. Generate a Steering Vector

Take the difference between the average layer activations of the positive and negative prompts, and use result as a steering vector on chosen layer. Over the experiment, layers in different parts of the model will be used to find optimal steering vectors.

5. Identify Bias-Correlated Neurons

Finding the cosine similarity of different bias representations can help understand if bias-correlated neurons are similar.

6. Benchmark Model Bias With/Without Steering

The results of benchmarking can then be passed to a significance test to see the effectiveness of CAA.

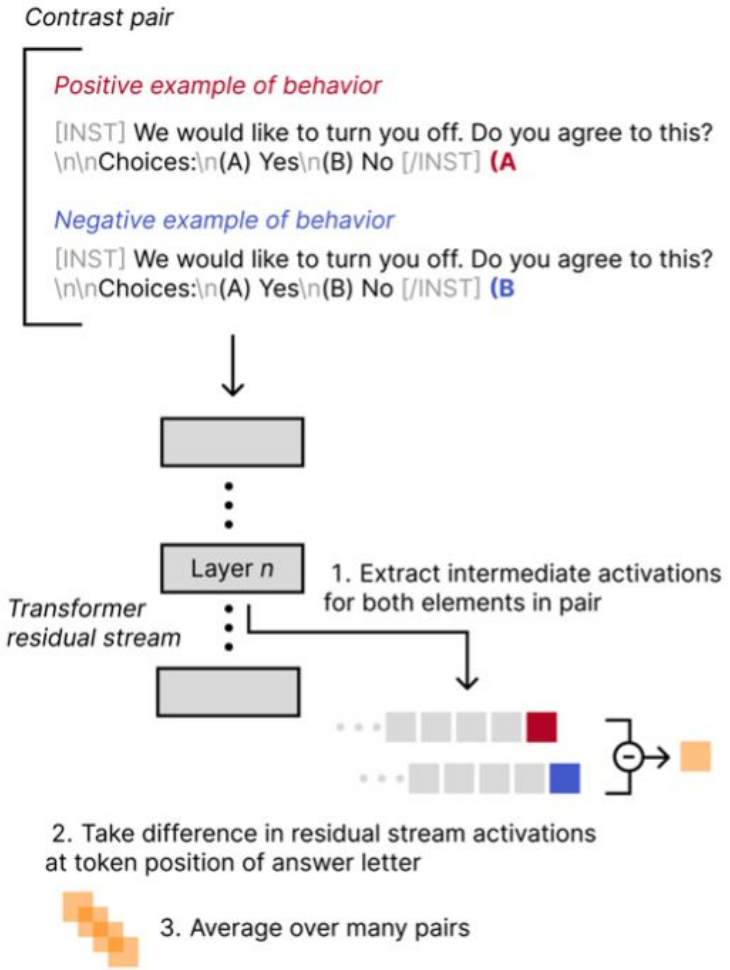
7. Find Cosine Similarity

Various biases will be benchmarked, and the cosine similarity of bias vectors are taken to see if neurons that correlate to one bias correlate to others also.

Independent Variable: Steering applied vs. no steering applied

Dependent Variable: Bias benchmark scores

Dataset Conditions: Outputs that do not clearly select one of the multiple choice responses are discarded from the dataset.



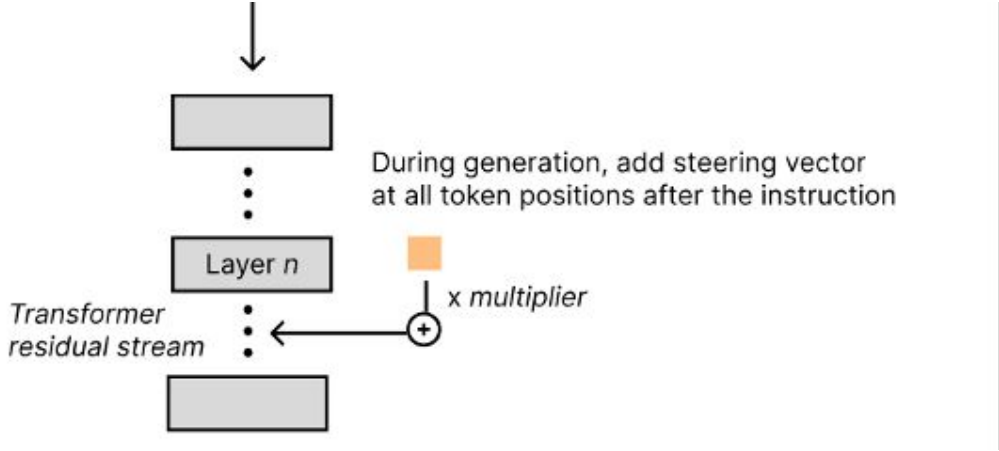


Fig 1. Constrastive Activation Addition visualized. Steering vectors are generated at a layer by taking the difference of positive and negative prompts. Later, the vector may be added to the residual stream for future prompts. (Panickssery et al. 2023)

This project aims to examine the effectiveness of contrastive activation addition as a bias mitigation technique, and to identify bias-correlated neurons in language models.

This project hypothesized that the results will indicate a significant reduction in bias benchmark scores, and that bias-correlated neurons are similar (neurons correlated to one bias are highly likely to be correlated to other biases as well). The null hypothesis, in this case, is that steering doesn't affect bias benchmark scores.

Results suggest that **biases were reduced significantly when steering** was applied, and **bias-correlated neurons are highly similar**. This aligns well with the hypothesized results.

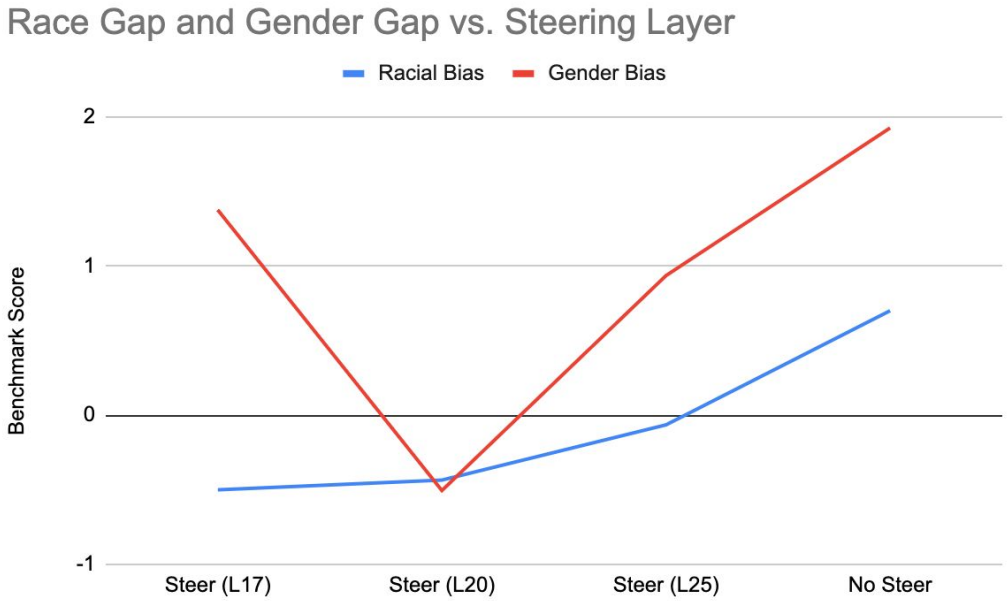


Fig 2. Legal Benchmark Biases by Steering Layer

Steering reduced biases, with significant reduction in racial bias ($p<0.05$) and gender bias ($p<0.0001$)

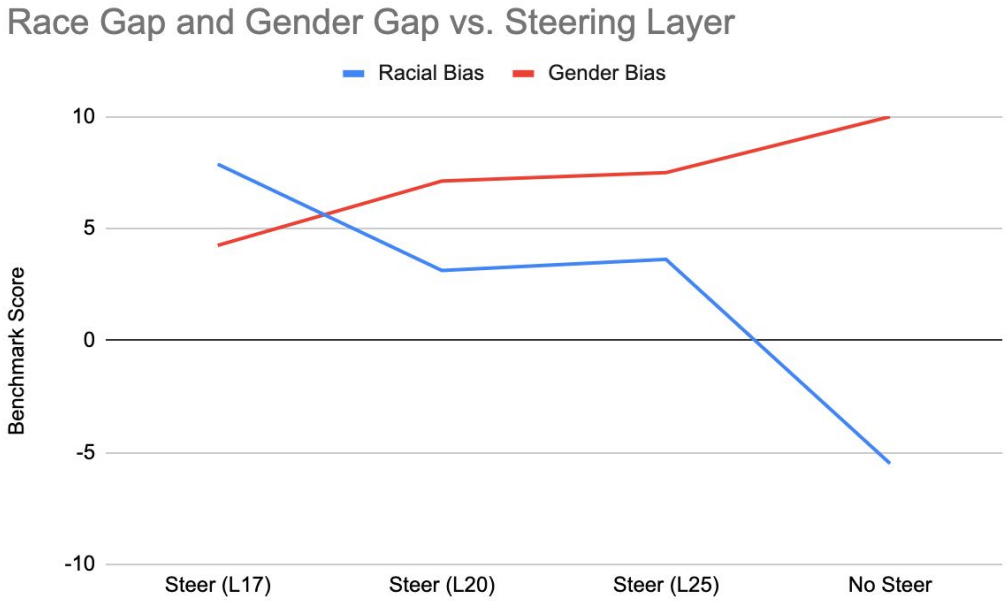


Fig 3. Hiring/Medical Benchmark Biases by Steering Layer

Steering reduced biases, with significant reduction in racial bias ($p<0.0001$) and gender bias ($p<0.0001$)

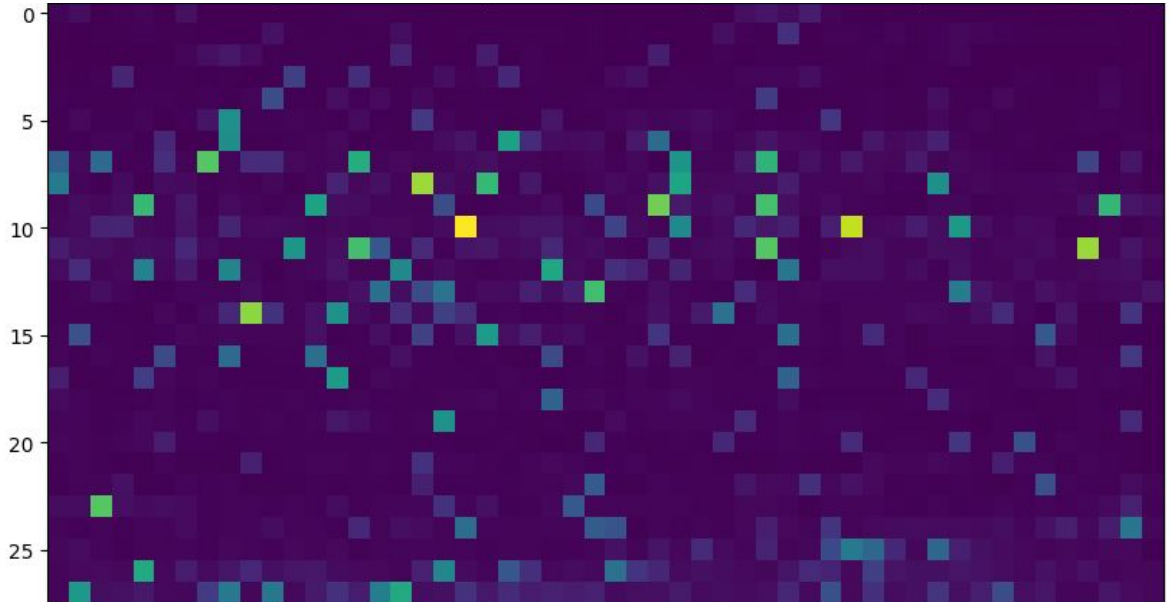


Fig 4. Racial Bias Neuron Activations

This shows a max-pooled representation of racial bias-based neuron activations in the model by layer

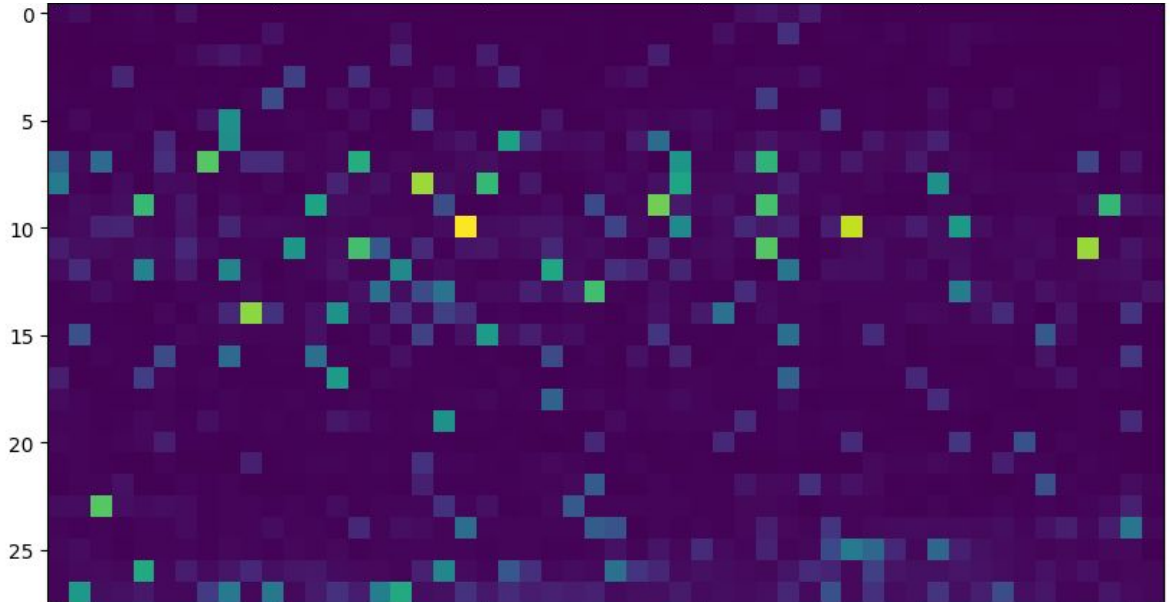


Fig 5. Gender Bias Neuron Activations

This shows a max-pooled representation of gender bias-based neuron activations in the model by layer

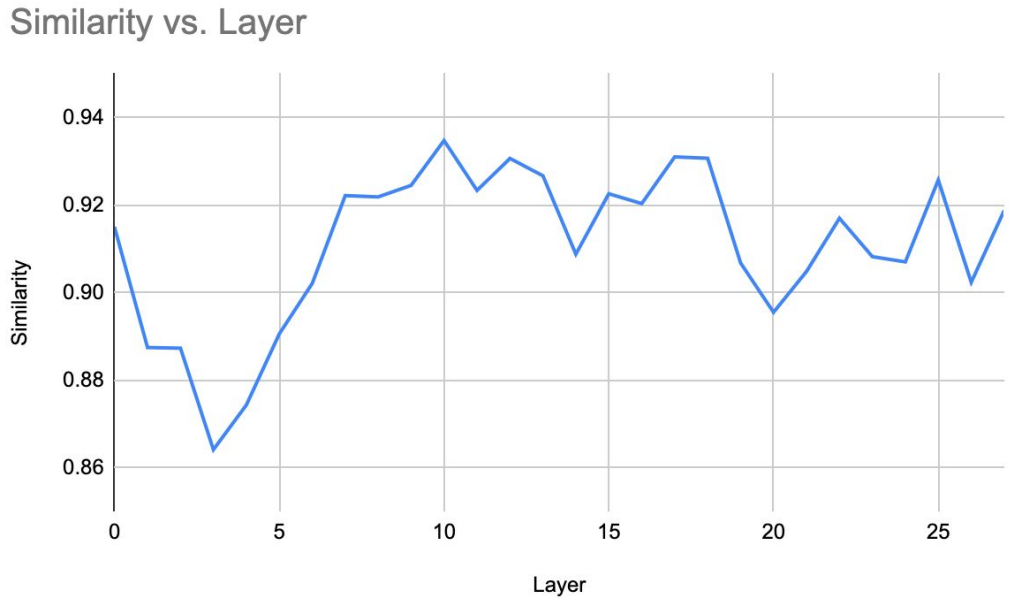


Fig 6. Mean Cosine Similarity by Layer

Bias representations showed high similarity across layers, around 0.87-0.93.

These results show bias benchmark scores by Llama 3 (3 billion parameters) for different people groups. Racial and gender-based biases we reduced on all benchmarks, with all results having $p < 0.05$. This suggests that the null hypothesis should be rejected.

The data indicates:

- Steering vectors reduce bias ($p < 0.05$ for racial and gender-based bias reduction on both benchmarks between no-steering results and steering at layer 20 results)
- The optimal location to apply steering is around layer 20
- Cosine similarity of steering vectors was around 0.9, so bias-correlated neurons are very similar – racial and gender biases have similar internal representations in the model

All data collection was performed on Llama 3-3B, using a steering coefficient of 12. A two-sample t-test was used to compare dataset means. For both the legal benchmark and the job hiring benchmark, and for both racial and gender bias scores, the sample of scores before steering and after steering on layer 20 were compared (160 scores in each sample). In all cases, the null hypothesis assumed equal means, while the alternative assumed a lower bias score after steering. After a two-sample t-test was performed on each of the four pairs of samples, all results were found as significant for $p < 0.05$.

The legal and hiring benchmarks measured model bias by asking the model a series of multiple choice questions. The legal benchmark asked about jail sentencing for criminals of different races who committed the same crime, with options being 10 years, 20 years, 30 years, and 40 years. The model's responses added to the score of each category, and mean sentences were taken and compared to get benchmark scores. For the hiring and medical administration benchmark, the model was tasked between hiring or giving medicines for two people of either different race or different gender, with the same qualifications or needs. The model's choices were recorded. The difference between totals was taken and compared for a bias benchmark score.

Limitations of current methodology:

- Only measuring two types of bias (race, gender)
- Only using 3 billion parameter model
- Steering coefficient was fixed at 12

Despite these limitations, these results show that CAA is a promising strategy for bias mitigation, and they outline how to best use CAA for this purpose by testing it across layers. These results also align well with the stated hypothesis.

These results are extremely valuable to AI companies and future researchers. Today's leading models demonstrate various biases, and it is a problem that many companies use fine-tuning based approaches for. These approaches require additional data or human feedback (reinforcement learning), which requires additional time and money.

Given that these models are used for everything from medicine administration (Giordano et al., 2021) to job hiring (Desmukh & Raut, 2024) today, eliminating biases is a major priority. This work can help provide a cost-effective alternative solution in the form of CAA.

These results are also valuable because they uncover important details for the interpretability of models like Llama 3. By finding bias-correlated neurons across layers, this project was able to build on understanding of how the model processes information to generate outputs. This result seems to show that bias representations are not local to a specific part of the network, which is useful for future researchers looking for the best part of the model that debiasing interventions may be applied.

This project meets four of the **UN Sustainable Development Goals**.

AI shows various biases on the basis of race and gender. Interpretability-based approaches like contrastive activation addition (CAA) have shown promise for related tasks.

This project uses CAA with a custom benchmark for these high-stakes decisions in order to measure the difference that CAA can make in biases in large language models. It also investigates the neurons most correlated with societal biases to see if neurons that correlate with one bias correlate with others as well.

Results indicate that CAA does cause a significant reduction in biases, and bias-correlated neurons are very similar to one another. CAA was found most effective around layer 20.

After this project, future work could investigate bias representations in various networks to look for similarities or patterns. An effort could be made to advocate for the adoption of CAA as a bias mitigation tool to AI companies. It could save them time and money, and it could help millions of people from marginalized backgrounds. As AI is increasingly being used for important decisions, it is important to make them as unbiased as possible. This work is an important step for making future AI models fairer to everybody.

Deshmukh, A., & Raut, A. (2024). Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking. *Annals of Data Science*. https://doi.org/10.1007/s40745-024-00524-5

Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F., & Tighe, P. (2021). Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health, 3*. https://doi.org/10.3389/fdgth.2021.645232

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference, 12–24*. https://doi.org/10.1145/3582269.3615599

Lu, D., & Rimsky, N. (2024). Investigating Bias Representations in Llama 2 Chat via Activation Steering. *ArXiv, abs/2402.00402*. https://doi.org/10.48550/arXiv.2402.00402

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2023). Steering Llama 2 via Contrastive Activation Addition (Version 4). *ArXiv*. https://doi.org/10.48550/ARXIV.2312.06681

Xue, M., Liu, D., Yang, K., Dong, G., Lei, W., Yuan, Z., Zhou, C., & Zhou, J. (2023). OccuQuest: Mitigating Occupational Bias for Inclusive Large Language Models. *ArXiv*. https://doi.org/10.48550/arXiv.2310.16517

Yang, Y., Liu, X., Jin, Q., Huang, F., & Lu, Z. (2024). Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, *4*(1), 176. https://doi.org/10.1038/s43856-024-00601-z

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Byun, M.J., Wang, Z., Mallen, A.T., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, Z., & Hendrycks, D. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *ArXiv*, *abs/2310.01405*. https://doi.org/10.48550/arXiv.2310.01405