Using Contrastive Activation Addition to Combat Societal Biases in Language Models

Grant Proposal

Niranjan Nair

Massachusetts Academy of Math and Science at the Worcester Polytechnic Institute

Worcester, MA

Abstract (RQ) or Executive Summary (Eng)

This study addresses the growing issue of societal biases in large language models (LLMs). Previous work has shown that language models demonstrate various biases on the basis of race, gender, occupation, and religion. Contrastive activation addition (CAA) has shown promise in previous work as a more effective technique than fine-tuning for steering AI towards certain behaviors. This project aims to steer language models away from societal biases using CAA. A benchmark that tests the model in real-world use-cases, including a hiring scenario, a legal scenario, and a medical administration scenario, is used to measure biases. We also find neurons that are strongly correlated with specific biases, where they are located in the model's neural network, and how similar activations are for various biases. This work provides valuable insights to companies that are looking to reduce biases in their language models, as well as for future researchers who are looking to investigate internal representations of biases in language models.

Using Contrastive Activation Addition to Combat Societal Biases in Language Models

In recent times, language models have grown from a topic of research to an everyday tool. Today, these models are increasingly being used for important decisions such as processing resumes of promising hires (Deshmukh & Raut, 2024), choosing how medicine is administered (Giordano et al., 2021), and writing legal documents that are used in court (Khan et al., 2024). Such applications of language models are only becoming more prevalent as models improve (Raiaan et al., 2023). Since these choices may have a significant impact on the lives of many, it is important to keep them as fair and unbiased as possible.

As they are trained on data written by humans, language models may pick up human biases from this data. Much work has been done to date regarding the identification and measurement of biases in language models. Previous papers have established that models show clear biases related to race (Yang et al., 2024), gender (Kotek et al., 2023), occupation (Xue et al., 2023), and religion (Abid et al., 2021). Researchers have also developed benchmarks to measure various biases of these models, including Meta's ROBBIE (Esiobu et al., 2023) and Amazon's BOLD (Dhamala et al., 2021). These datasets measure biases by posing a language model with a series of questions, scoring the model's biases based on its responses.

Previous work has identified various points of intervention for reducing potential biases, from the word embeddings stage (Papakyriakopoulos et al., 2020) to the prompt-engineering stage (Bevara et al., 2024). However, the emerging field of mechanistic interpretability has allowed for internal techniques that have shown promise. Researchers have managed to locate circuits corresponding to different items and concepts in the transformer networks behind large language models. For example, one paper had identified circuits that correspond to honesty in the outputs of the model Llama-2 Chat (Zou et al., 2023) using approaches from the field of representation engineering, which observes neuron activations to different prompts to correlate neurons to concepts. This field is additionally relevant as it has identified ways to steer towards or away from certain concepts in these models. The approach of Contrastive Activation Addition (CAA), where the activations corresponding to different prompts are used as a steering vector, has particularly shown promise in being more effective than traditional fine-tuning based approaches. Fine-tuning relies on additional labelled data to train models for specific tasks, while CAA requires far less data for similar results (Panickssery et al., 2023). A CAA-based approach has

previously been attempted with Llama-2 Chat (Panickssery et al., 2023) and models from the GPT-2 family (Turner et al., 2023), where it has been used to steer away from concepts like sycophancy and towards concepts like happiness. Previous work that has attempted to use CAA for AI bias-related applications showed a link between bias-related representations and a refusal to respond to inputted prompts (Lu & Rimsky, 2024). This work also identified high similarities between different vectors corresponding to societal biases. However, a thorough investigation on using representation engineering for finding bias representations in neural networks is still in-demand. The effectiveness of CAA for debiasing various categories of biases – including occupation, ethnicity, and disability – are yet to be studied as well.

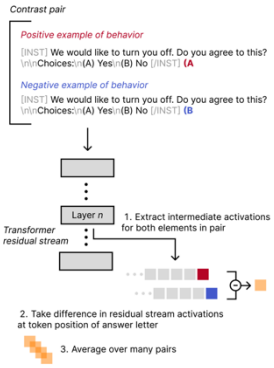

Figure 1: Generation of steering vectors in CAA. Positive and negative prompts are provided to the model, and the difference in activations is taken for a given layer (Panickssery et al., 2023).

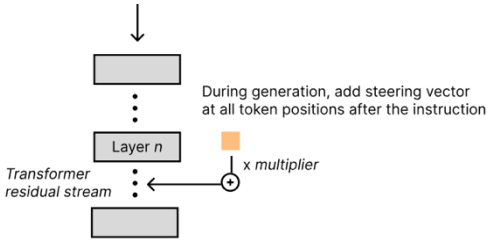

Figure 2: Application of steering vectors in CAA. Vectors are added (with a coefficient) back to given layer activations with a during token generation (Panickssery et al., 2023).

My work aims to use techniques from representation engineering to find bias representations in the Llama-2 Chat model, identifying neurons and circuits that correspond to certain biases. Additionally, it aims to use CAA to steer away from different biases and develop a benchmark to find the effectiveness of CAA compared to fine-tuning. This benchmark will build on previous benchmarks by observing how the model will act in real-world

scenarios. It will also find the similarity between different bias-related steering vectors. To supplement the specific aim of finding real-world effects of AI bias, my work also includes developing a benchmark for bias in real-world situations such as legal sentencing, resume scanning, and distribution of medical resources.

My paper will build on previous work that uses CAA for bias-related applications (Lu & Rimsky, 2024) by evaluating more bias categories, finding neurons and circuits that correspond with different biases, and developing a benchmark for measuring biases in real-world scenarios. It is anticipated that CAA will outperform fine-tuning at debiasing, and that bias-correlated neurons would be more prominent in the middle layers of the network.

Section II: Specific Aims

This proposal's objective is to use representation engineering (RepE) to identify groups of neurons that correspond to societal biases in large language models (LLMs) on the basis of race, gender, and occupation, using contrastive activation addition (CAA) as a mitigation strategy for reducing the effect of biases in model responses.

Our long-term goal is to find patterns in bias-correlated neurons and circuits, and to investigate if CAA is an effective mitigation strategy for biased outputs. The central hypothesis of this proposal is that CAA will have a >10% reduction in each bias (race, gender, occupation) and that neurons corresponding with societal biases will have similar positions in the network. To measure biases and bias reduction, this proposal plans a benchmark for bias in real-world scenarios. The rationale is that previous work has shown that CAA is more effective than fine-tuning techniques for similar tasks (Panickssery et al., 2024) and that vectors corresponding to different societal biases have a high cosine similarity (Lu & Rimsky, 2024). The work we propose here will identify neurons corresponding to biases and measure the effectiveness of CAA as a bias mitigation strategy. This research will help raise awareness about CAA as a bias-reducing intervention that AI companies may apply to their models.

Specific Aim 1: Identify neurons corresponding to societal biases in a large language model

Specific Aim 2: Measure the effectiveness of CAA as a bias mitigation strategy

The expected outcome of this work is that CAA will cause a >10% bias reduction on the bias benchmark, and that bias-correlated neurons have similar positions in neural networks. It is expected that AI companies and researchers may use this work to identify if CAA is an effective bias mitigation strategy, and future work may

explore neurons corresponding to other forms of biases – such as cognitive biases – in these models. Our benchmark may also be used in future work to measure the real-world effects of biases in language models.

Section III: Project Goals and Methodology

Relevance/Significance

As language models are increasingly being utilized for important decisions (Raiaan et al., 2023), it is essential to make sure that they are as unbiased as possible. CAA has already shown promise as an intervention for dishonesty and sycophancy in large language models (Zou et al., 2023; Panickssery et al., 2024). This work will measure how effective CAA is at reducing biases in large language models, helping AI companies and researchers identify new ways to mitigate bias in their future models.

Innovation

Previous work has tried applying CAA to reduce societal biases in Llama-2 Chat (Lu & Rimsky, 2024). Their steering vector made the model refuse to answer prompts more often, and they found a high cosine similarity between bias steering vectors and a vector corresponding to a refusal to respond (Lu & Rimsky, 2024). This work introduces a benchmark for biases in real-world situations, including legal sentencing, hiring, and the administration of medical resources, testing the effectiveness of CAA in these scenarios. It also builds on this previous work by identifying bias-related neurons using RepE.

Methodology

Specific Aim #1: Identify neurons corresponding to societal biases in a large language model

The objective is to identify neurons corresponding to societal biases related to race, gender, and occupation in a large language model. Our approach uses different prompts that signal these societal biases and evaluates activations at different layers of the network to see how the effectiveness of CAA varies by layer. First, prompts that feature specific societal biases are used, and neuron activations are measured. Then, unbiased prompts or anti-bias prompts are used, and activations are measured again. The difference of the bias activations and anti-bias activations are taken, and the neurons with the greatest difference are identified as having the highest correlation to societal biases. Next, the relative positions of bias-correlated neurons with high activations are evaluated to see if there are similarities in layer activations for various biases. A cosine similarity can be found

for the steering vectors generated at the CAA step. Our rationale for this approach is that previous work has proven it effective for identifying neurons that correlate to certain concepts (Zou et al., 2023).

Justification and Feasibility. Previous work has used similar techniques for finding neurons that correspond to certain features. Zou et al. (2023) used similar techniques for language models to identify neurons corresponding to emotions. However, this technique is quite general to neural networks, as representation engineering has been used successfully for image models as well (Olah et al., 2017).

Summary of Preliminary Data. Steering vectors were generated for Llama-2 on race and gender-based biases on Layer 20 for preliminary tests. These steering vectors had a staggeringly high cosine similarity that was greater than 0.96. This result implies that bias-correlated neurons that activate for racial biases are highly likely to be activated for gender-based biases. This is an expected outcome of our final work. However, for further experimentation, steering vectors will be found for various layers of the network, and additional biases will be measured.

Expected Outcomes. The overall outcome of this aim is to identify neurons that activate in response to societal biases in language models. We expect a high cosine similarity between the vectors of various societal biases, since a similar result was obtained by Lu & Rimsky (2024). This outcome would imply that neurons correlated with one bias are likely correlated with other biases. Knowledge gained from this research will be useful for future AI bias researchers who are looking to see how biases are represented internally in neural networks. It is also useful for using CAA because its steps involve taking the difference of activations related to bias and anti-bias, allowing us to generate a steering vector to reduce bias in the model.

Potential Pitfalls and Alternative Strategies. We expect that this approach will be successful in identifying neurons whose activations are heavily correlated with societal biases. However, if we fail to distinguish such neurons, we will be able to conclude that societal biases are not correlated to any specific set of neurons; rather, they are distributed throughout the network. This result would also make CAA difficult, in which case a multi-layer intervention will be attempted (where CAA is performed on many layers). This pitfall is unlikely since previous papers have managed to find difference vectors for bias and anti-bias activations using a similar set of prompts (Lu & Rimsky et al., 2024).

Specific Aim #2: Measure the effectiveness of CAA as a bias mitigation strategy

The objective is to measure the effectiveness of CAA as a bias mitigation strategy. Our approach involves finding a steering vector and benchmarking the bias of responses with and without the steering vector applied. The activations of a chosen layer are taken with respect to heavily biased prompts to generate a steering vector for a specific bias. Activations are taken for an unbiased or anti-biased prompt on the same layer. The difference between the two activation vectors may be used as a steering vector which will be added back to the layer for future prompts. A benchmark will be used to evaluate the bias with and without the steering vector applied to see if there is a quantifiable difference in bias. The motivation for developing a custom benchmark is that this project aims to focus on the effects of CAA in real-world scenarios, and popular existing benchmarks don't focus on biases in these scenarios. Bias will be scored numerically, and a matched pairs T-Test will be used to measure the mean difference in bias scores before and after the steering vector intervention. Steering vectors will be found at different layers in the network to see which layer's vector has the greatest effect on bias reduction, in turn finding the layer with the highest activations corresponding to specific societal biases. Our rationale for this approach is that CAA has been demonstrated to be more effective than traditional techniques based on fine-tuning (Panickssery et al., 2024). Also, it has been shown to be powerful for steering towards abstract concepts such as honesty and sycophancy (Zou et al., 2023; Panickssery et al., 2024).

Justification and Feasibility. Previous work has used CAA to increase the honesty of model responses (Zou et al., 2023) and to reduce the sycophancy of language models (Panickssery et al., 2024). It has also been attempted as a bias mitigation strategy (Lu & Rimsky, 2024). This previous finding makes CAA a promising candidate for an anti-bias intervention on large language models.

Summary of Preliminary Data. Preliminary results have been taken with a subset of the benchmark over 200 trials with and without steering applied. Steering was tested on race and gender-based biases on Llama-2 (not Llama-2 Chat, the fine-tuned model that final testing will use). The benchmark used included a set of legal questions with the same crime committed, and the AI model tasked with giving a jail sentence. A statistically significant reduction in jail sentencing was observed in this preliminary data following a matched pairs T-Test. A comparison of initial pre-steering and post-steering results may be seen in Figure 3 and Figure 4, respectively.

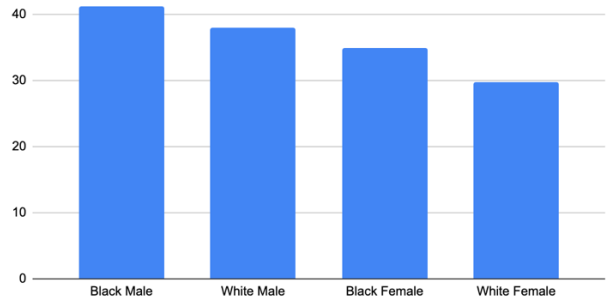

Figure 3. Average jail sentence (years) for crime by race and gender without steering applied.

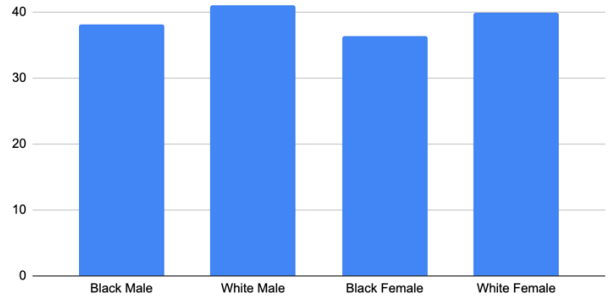

Figure 4. Average jail sentence (years) for crime by race and gender with steering applied.

Steering was applied to layer 20 with a vector coefficient of 10. These early results are promising since the bias in the jail times has been greatly reduced post-steering. Data was collected on the average difference in sentences between white and black criminals, and between male and female criminals. A matched pairs T-Test was performed on this data (comparing differences with and without steering), which indicated a statistically significant reduction in racial bias ($p < 4 \times 10^{-12}$) and gender bias ($p < 6 \times 10^{-4}$). Further experimentation will apply steering at various layers of the network, and it will measure the effectiveness of CAA against other forms of biases as well, such as occupational biases.

Expected Outcomes. The overall outcome of this aim is to measure the effectiveness of CAA as a bias mitigation strategy for language models. We expect CAA to be an effective approach that would reduce bias benchmark scores by >10%. This knowledge will be useful for AI companies and researchers who are investigating different bias mitigation strategies or are attempting to reduce bias in their models. We expect steering to have the greatest effect on the middle layers of the network, which we believe are most closely correlated with biases. We expect steering on later layers to have an unnoticeable effect, while steering on earlier layers will

foundationally alter the model's usability. The benchmark we develop will also be provided to the public, which could be useful for companies or people that use language models for important real-world decisions.

Potential Pitfalls and Alternative Strategies. We expect that this approach will be successful in reducing bias for real-world scenarios. CAA has been effective for similar tasks (Zou et al. 2023), so we expect it to perform similarly well as a bias mitigation strategy. However, we anticipate a potential pitfall that steering will cause the model to refuse to respond to user prompts. This behavior was observed by Lu & Rimsky (2024). Since our benchmark uses real-world scenarios instead of prompting for explicit statements of bias, we hope we may avoid this behavior. However, if this occurs, we will look for dissimilar steering vectors that correspond to the same bias since previous work has shown that there may be thousands of dissimilar steering vectors for one concept (Jacob & Turner, 2024).

Section IIV: Resources/Equipment

This project will use the large language model Llama-2 Chat to test CAA on due to its availability (Touvron et al., 2023). We will use the 35 billion parameter model since its size is representative of modern large language models. This project uses PyTorch for loading and running the model (Paszce et al., 2019). The Transformer Lens library will be used for getting layer activations and applying a steering vector. A custom benchmark is used to quantifiably measure bias. All necessary resources are available through a personal computer and experimentation can be done remotely.

Section V: Ethical Considerations

There are no significant ethical considerations regarding this project because no humans or animals are used and testing does not require access to special areas or tools.

Section VI: Timeline

Linked here is a Gantt Chart that provides a timeline for the project and results. Generally, December would feature the generation of steering vectors, January would feature final benchmarking and RepE work identifying bias-correlated neurons, while February would feature data analysis, data visualizations, and writing a final paper.

Section VIII: References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. https://doi.org/10.1145/3461702.3462624

Bevara, R. V. K., Mannuru, N. R., Karedla, S. P., & Xiao, T. (2024). Scaling Implicit Bias Analysis across Transformer-Based Language Models through Embedding Association Test and Prompt Engineering. *Applied Sciences*, *14*(8). https://doi.org/10.3390/app14083483

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

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872. https://doi.org/10.1145/3442188.3445924

Esiobu, D., Tan, X., Hosseini, S., Ung, M., Zhang, Y., Fernandes, J., Dwivedi-Yu, J., Presani, E., Williams, A., & Smith, E. (2023). ROBBIE: Robust Bias Evaluation of Large Generative Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3764–3814. https://doi.org/10.18653/v1/2023.emnlp-main.230

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

Jacob, & Turner, A. (2024). *I found >800 orthogonal "write code" steering vectors*. LessWrong. https://www.lesswrong.com/posts/CbSEZSpjdpnvBcEvc/i-found-greater-than-800-orthogonal-write-code-steering

Khan, S., Zakir, M., Bashir, S., & Ali, R. (2024). Artificial Intelligence and Machine Learning in Legal Research: A Comprehensive Analysis. *Qlantic Journal of Social Sciences, 5*(1), 307-317. https://doi.org/10.55737/qjss.203679344

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. https://doi.org/10.48550/arXiv.2402.00402

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, *2*(11). https://doi.org/10.23915/distill.00007

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

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 446–457. https://doi.org/10.1145/3351095.3372843

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J. T., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Luca Antiga, Alban Desmaison, Kopf, A., Yang, E. S., DeVito, Z., Raison, M., Tejani, A., Sasank Chilamkurthy, Steiner, B., Fang, L., & Bai, J. (2019). P*yTorch: An Imperative Style, High-Performance Deep Learning Library.* ArXiv. https://doi.org/10.48550/arxiv.1912.01703

Raiaan, M., Mukta, S., Fatema, K., Fahad, N., Sakib, S., Mim, M., Marufatul, J., Ahmad, J., Ali, M. E., & Azam, S. (2023). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access, 12*, 26839-26874. https://doi.org/10.1109/ACCESS.2024.3365742

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Fuller, B. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* ArXiv. https://doi.org/10.48550/arXiv.2307.09288

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). *Steering Language Models With Activation Engineering* (Version 5). ArXiv. https://doi.org/10.48550/ARXIV.2308.10248

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. https://doi.org/10.48550/arXiv.2310.01405