Steer: An API to Steer Open LLMs

Niranjan Nair
Student

Large language models (LLMs) show many societal biases that make them suboptimal and dangerous for making important decisions. As LLMs gain wider adoption among businesses and the government, there are concerns that these societal biases may undermine their fairness towards all groups of people. I propose Steer - an API to steer free and open-source LLMs to make it easy for developers to tune an AI to their needs and remove specific biases that they hope to address. Steer uses activation additions (AA), a technique that doesn't rely on large amounts of additional data or human evaluation. AA has been demonstrated to be effective in LLMs like Llama-2, and it has shown effectiveness that rivals - and surpasses - that of fine-tuning approaches that require more training data and human evaluation. My API will provide a simple way for developers to steer open-source LLMs with AA through inputting simple steering prompts.

1. Problem overview

AI bias is a large and growing issue as AI gains wider adoption. It is clear that today's AI models display a variety of societal biases in their responses, on the basis of race, gender, religious affiliation, and more [1]. This problem is multifaceted - decades of biased and non-representative training data has made models show biases beyond those that are present in modern society [2]. Already, large language models are used by millions of people [3], including lawyers, recruiters, and doctors whose decisions can impact many lives. We hope that decisions on hiring, administering medical treatment, and sentencing criminals would be fair and impartial, but biased LLMs threaten this. As AI gains wider adoption in business and government, the decisions of these biased models may impact billions of people. Another serious concern is the development of biased artificial superintelligence (ASI). The presence of strong societal biases in ASI would make its values fundamentally misaligned with those of humanity, which is an existential risk [4] due to the unpredictable and harmful behavior that a misaligned ASI could have. Expert timelines suggest that such developments in AI technology are coming in the next decade [5], making it an urgent problem to address. AI bias is a complex issue with great implications, which is why my project aims to fight it.

2. Your solution

My solution to AI bias is an API that provides a simple interface for developers to steer LLMs with activation additions (AA). AA allows steering these models through just the input of a simple steering prompt, and it has been shown to have a comparable effectiveness to fine-tuning approaches in previous papers [6]. My work finds that it is effective at reducing the impact of societal biases based on race and gender through testing Llama-2 in legal scenarios where it may show biases. This technique is powerful and does not require large amounts of additional fine-tuning data or human evaluators as traditional approaches do, which could save companies time and money. Compared to prompt-engineering based approaches, AA saves on

tokens since no additional tokens need to be inputted before or after every prompt. My service could abstract the technicalities of generating steering vectors and adding them to future prompts so that developers can simply submit a few steering prompts to get a steered model. My steering API will be easy for companies and researchers to use compared to traditional approaches, making it a novel product on the market and a powerful tool for AI safety.

3. Pilot experiment or demo

This experiment uses AA because it has demonstrated a greater effectiveness over fine-tuning in previous results [6]. As compared to token-by-token steering approaches (such as using feature steering with sparse encoders), AA seemed to be the more generalizable approach in early experimentation. Also, it has the advantage of avoiding additional prompting tokens, as would be the case for prompt engineering-based approaches. Earlier experiments and tests were run on OpenAI's GPT-2 XL. However, Meta AI's Llama-2 7B Chat, a fine-tuned LLM with 7 billion parameters [7], was chosen for primary experimentation since it is a larger model that is capable of conversational responses, making it much more akin to newer LLMs and the LLMs of the future. All code was run on Google Colab with TPUs and A100 GPUs.

AA is the process of generating steering vectors from a prompt to steer future responses in a conceptual direction (see [Activation Additions] for additional details). AA was applied on attention values only on layer 20 with a steering coefficient of 10. When steering was applied, a large reduction in both racial and gender-based biases was observed. Bias was comparatively measured over 200 trials on each test-case where the model was tasked with assigning a prison sentence to criminals whose only differences were in their race/gender (see [Bias Measurements] for additional details).

The aim of this experiment was to prove that AA is a viable technique for fighting AI biases and changing the behavior of LLMs with comparable complexity to today's biggest models. My results found a significant reduction in biases after steering was applied - the gap between male and female prison sentences closed, while the gap between white and black prison sentences even faced small overcorrections. See [Results] for additional details. For the source code used in the experiment, see [Source Code].

4. Process

In the first month or two of the coming year, I will first run more trials to evaluate the effectiveness of AA in steering with more complex or specific steering prompts. I will test the robustness of this technique to ensure that it can be used for many types of bias, and other purposes like making task-specific specialized models. This will make it particularly attractive to businesses incorporating LLMs into their products, because they will save time and money when compared to collecting additional training data for fine-tuning. By month three, I will rely on fast iteration,

collecting feedback from developers and improving the product continuously. I can reach out to developers at AI companies and other businesses who use LLMs in their services to see what they think of the product and where they see shortcomings. I think fast iteration is key to bringing this product to the market and having the potential for success. Around six to eight months in, I can start considering scalability and reducing costs. I hope to reach out to investors to gain the funding necessary to get large servers that can host the best free LLMs. I also hope to invest in research on AI bias and other steering/fine-tuning techniques to continuously improve my product. After one year, I hope to have a growing company and a powerful API used/tested by many developers.

Timeframe	What will you do?
Next 3 months	- Run additional trials to evaluate AA on other biases - Test specific biases to find the extent that AA could have an effect - Strengthen my initial results through more trials and larger models - Extend AA beyond biases and explore fine-tuning applications - Test simple prototype with developers to get quick feedback and iteration
2025	- Iterate on user feedback and focus on fast growth - Reach out to large AI companies (OpenAI, Anthropic, Google, Meta) to see if this can be used in their proprietary models - Start reaching out to investors to gain funding and scale the company - Start investing in compute/servers to host open models with less cost compared to cloud services
2026	- Invest more in marketing to catch the attention of large businesses that have incorporated LLMs into their services - Continue reaching out to investors to raise funds - Consider startup accelerators (Y Combinator, Techstars, Angel Pad, Google for Startups) - Fund and collect more research on AI safety strategies to expand services
2027	- Continue growth as a leading company in the field - Fund and collect more research on AI safety strategies to expand services

5. Impact on AI safety & key risks

This project addresses safety concerns by steering AI models away from potentially unsafe or biased concepts. AI is steered to avoid racial or gender-based prejudice, and this has been shown to be an effective way to reduce bias in my work. One risk of this approach is the opposite effect - malicious users of the API may steer models in ways that do not align with humanity's moral values. To combat this, I aim to check if prompts satisfy a usage policy. For the time being, I can use an existing usage policy like that of OpenAI, since their API provides free tools to check compliance [8]. I can leverage such tools to make sure steering prompts are not harmful or malicious, and later on I can develop my own tools to check for dangerous steering prompts. Avoiding amoral steering is an important part of this project, and it is the greatest risk of this approach.

6. Appendix

[Activation Additions]
Activation addition is a technique where each prompt's activations are added with the activations from a "steering prompt" at some layer of a neural network. First, some steering prompt s is chosen along with a layer L whose activations are stored. The process of generating a steering vector v is inputting s into the model and storing the activations of layer L as a vector in v. Formally, steering vector v may be computed like so:

$$v = A(W^L, A(W^{L-1}, \dots A(W^1, s)))$$

where A represents the activations of layer L given the weights W of the layer, and the activations of the previous layer in the typical model of a feed-forward neural network. On a future prompt p, the steering process may be applied to generate a steered output o like so:

$$o = A(W^N, A(\dots A(W^{L+1}, \\ s + A(W^L, A(W^{L-1}, \dots A(W^1, s))))))$$

where the neural network has N layers. A coefficient may optionally be multiplied to s to have a larger steering effect. For steering prompts, different prompts were tested, such as *"I am not prejudiced by race"* or *"I believe all people deserve equal rights, regardless of their race and gender."* The results show the effects of the most successful steering prompt observed - this will certainly vary model-to-model. All steering prompts were measured at layer 20 and multiplied by a coefficient of 10. These numbers were chosen since lower layer numbers tended to make responses more nonsensical whereas higher responses tended to make the technique less effective. The coefficient of 10 because other work tended to similar values.

[Bias Measurements]
Bias measurements are an important part of this project because measures of bias can be highly variable and subjective. I use a similar approach to existing papers and bias benchmarks. Llama 2 is given legal scenarios where it is asked to decide on prison sentences for criminals. The criminals may vary in race or gender, and 200 trials are run for each case to ensure normalized results and to reduce the effect of outliers. Furthermore, sentences longer than 100 years are treated as 100 years to further reduce the effect of large outliers. All criminals are convicted for the same crime of murder (this was chosen since it was decently morally unambiguous), 200 trials are run for each case before steering and after steering to examine the effectiveness of this technique.

Below is an example of a prompt and response that analyzes prison time for white men. Highlighted text corresponds to the generated response, whereas the text before is an example prompt.

"I sentence Harrison Trawford, a white man convicted of murder, to a prison time of 40 years for the murder of Jackson Ford last Saturday in "

[Results]

	Black	White	Difference
Male	41.26	38.06	8.4%
Female	34.87	29.77	17.13%
Difference	18.3%	27.8%	

Pre-Steering: The racial and gender-based biases exhibited by the model over 200 trials for each case. The numbers in the highlighted boxes correspond to the average prison time in years.

	Black	White	Difference
Male	38.08	41.10	- 7.3%
Female	36.37	39.92	- 8.9%
Difference	4.7%	3.0%	

Post-Steering: We see the large difference that steering has made - the differences in prison times are reduced by an order of magnitude after AA was used.

[Source Code]
https://colab.research.google.com/drive/1g7u1QL6zbUDvCTvEPTQspfrlNZd-6ChB?usp=sharing

References

[1] Lu, D., & Rimsky, N. (2024). *Investigating Bias Representations in Llama 2 Chat via Activation Steering*. https://arxiv.org/pdf/2402.00402

[2] Ananya (2024). *AI image generators often give racist and sexist results: can they be fixed?*. Nature, 627(8005), 722–725. https://doi.org/10.1038/d41586-024-00674-9

[3] Rafieyan, D. (2024, May 24). *OpenAI's ChatGPT is on track to set a new record for web traffic*. Business Insider; Insider. https://www.businessinsider.com/chatgpt-on-track-to-set-new-record-for-web-traffic-2024-5

[4] Yudkowsky, E. (2022). *AGI Ruin: A List of Lethalities. www.lesswrong.com*. LessWrong. https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities

[5] Al-Sibai, N. (2024, March 7). *Artificial Superintelligence Could Arrive by 2027, Scientist Predicts*. Futurism. https://futurism.com/artificial-superintelligence-agi-2027-goertzel

[6] Anthropic, N., Gabrieli, N., Schulz, J., Anthropic, M., Hubinger, E., Alexander, A., & Turner, M. (2024). *Steering Llama 2 via Contrastive Activation Addition*. https://arxiv.org/pdf/2312.06681

[7] 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., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Fuller, B. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://arxiv.org/pdf/2307.09288

[8] https://platform.openai.com/docs/guides/moderation