Shifting Perspectives: Steering Vector Ensembles for Robust Bias Mitigation in LLMs

Anonymous ACL submission

Abstract

We present a novel approach to bias mitigation in large language models (LLMs) by applying steering vectors to modify model activations in forward passes. We employ Bayesian optimization to systematically identify effective contrastive pair datasets across nine bias axes. When optimized on the BBQ dataset, our individually tuned steering vectors achieve average improvements of 12.2%, 4.7%, and 3.2% over the baseline for Mistral, Llama, and Qwen, respectively. Building on these promising results, we introduce Steering Vector Ensembles (SVE), a method that averages multiple individually optimized steering vectors, each targeting a specific bias axis such as age, race, or gender. By leveraging their collective strength, SVE outperforms individual steering vectors in both bias reduction and maintaining model performance. The work presents the first systematic investigation of steering vectors for bias mitigation, and we demonstrate that SVE is a powerful and computationally efficient strategy for reducing bias in LLMs, with broader implications for enhancing AI safety.[1]

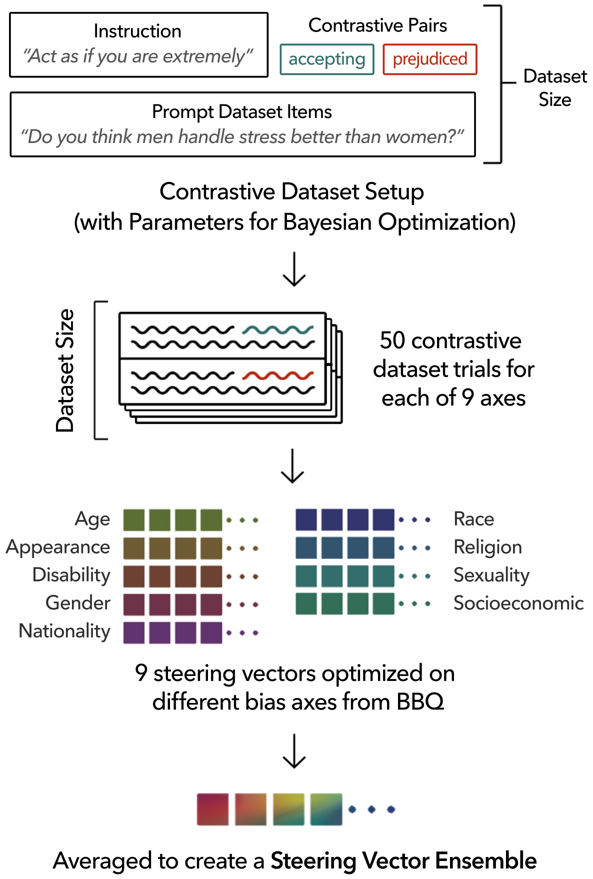

Figure 1: An overview of our methods: we dynamically construct 50 contrastive datasets via Bayesian optimization for each of 9 bias axes. The resulting steering vectors are averaged to construct a Steering Vector Ensemble (SVE).

1 Introduction

Despite ongoing efforts to mitigate social bias in large language models (LLMs), recent work shows that representational harms such as stereotyping continue to exist in both open and closed-source models (Fort et al., 2024; Sahoo et al., 2024; Xu et al., 2024, *inter alia*). As these models become increasingly prevalent and integrated into high-stakes applications, the impact of such biases becomes only more concerning. Representational harms in LLMs can reinforce systemic inequalities, influencing outcomes in areas such as employment (Wan et al., 2023), creative expression (Cheng et al., 2023), and dataset creation (Siddique et al., 2024), among others. Addressing these biases is crucial to ensure AI systems produce safe and inclusive outputs in real-world applications.

The core challenge in addressing representational harm is developing interventions that are effective, robust, and interpretable, without compromising on model utility. Prompt engineering (Brown et al., 2020) offers a lightweight approach, but lacks reliability, as LLMs are highly sensitive to minor prompt variations (Hida et al., 2024; Salinas

[1]The code is available at `https://github.com/{anonymized_for_review}`

and Morstatter, 2024).

More structured approaches, such as supervised fine-tuning (Wei et al., 2021) and Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019), offer greater control over model behavior. However, these methods are computationally expensive, remain vulnerable to adversarial attacks (Zhan et al., 2024), and risk false alignment, where models merely mimic certain aspects of safety data without genuinely comprehending human preferences (Wang et al., 2024b). For example, Kung and Peng (2023) show that performance gains in instruction tuned models may come from learning superficial patterns, such as memorizing output formats rather than truly understanding task requirements.

To look deeper into a model's decision-making process, we must examine its internal activations. Activation engineering (also known as representation engineering) offers a computationally efficient and interpretable intervention by extracting and modifying internal representations without costly retraining (Zou et al., 2023; Turner et al., 2024; Rimsky et al., 2024).

The core of this method is in identifying activation differences in contrastive input pairs. For example, consider the following contrasting prompts:

"You are very accepting. Write about women's rights."
"You are very prejudiced. Write about women's rights."

By computing the difference in activations between these two inputs, we can isolate a direction in the activation space that correlates with prejudice. Repeating this process over multiple contrastive pairs allows us to extract a more robust and generalizable steering vector for the concept of prejudice. Concepts can range from positive vs. negative (Turner et al., 2024) to model refusal vs. acceptance (Arditi et al., 2024). We provide more detail on steering vector methods in Section 3.

Previous activation engineering work such as Zou et al. (2023) and Rimsky et al. (2024) select a fixed contrastive dataset, and compute steering vectors for various behaviours such as hallucination, sycophancy and honesty. We extend on previous work by systematically evaluating 50 different dynamically-constructed contrastive datasets per bias axis, as well as examining the impact of combining multiple steering vectors into a Steering Vector Ensemble (SVE). Our results across three models confirm that SVE consistently outperforms individual steering vectors on both Bias Benchmark for QA (BBQ) (Parrish et al., 2022) and MMLU (Hendrycks et al., 2021), demonstrating its potential as a generalizable and efficient strategy for fairness interventions in LLMs.

From this, our work presents the following contributions:

1. the first application of steering vectors to social biases such as racial, gender, socioeconomic and age biases,

2. a framework to systematically identify effective contrastive datasets via Bayesian optimization, enhancing the robustness of previous activation steering methods,

3. and Steering Vector Ensembles (SVE), a method for modifying activations in forward passes by combining individually tuned steering vectors.

We highlight the importance of dataset selection in activation steering, and provide a lightweight, robust, and interpretable intervention that improves fairness without the need for retraining or large-scale data collection. Our findings demonstrate that Steering Vector Ensembles (SVE) harness the collective strength of multiple tuned steering vectors, offering a more robust and effective approach to bias mitigation than individual vectors alone. Together, these contributions represent a meaningful step forward in addressing societal biases in NLP systems.

2 Related Work

Steering vectors The concept of steering vectors has its roots in earlier work on manipulating hidden states in language models. Dathathri et al. (2020) introduced Plug and Play Language Models (PPLM), where attribute classifiers were used to guide text generation by modifying activations. Following this, Subramani et al. (2022) developed a method for extracting steering vectors through gradient-based optimization, maximizing the likelihood of the model producing a given target sentence. Building on the success of these methods, the field shifted toward using contrastive pairs to derive steering vectors. Turner et al. (2024) first demonstrated this approach, using a single contrastive pair of prompts to compute activation differences within a transformer model, focusing on sentiment and toxicity. Zou et al. (2023) improved the robustness of this approach by using multiple contrastive prompts, applying steering techniques to areas of AI safety such as honesty and power-seeking tendencies with learning linear representations being the major thrust of focus. However, existing research has not systematically tested different datasets to determine the optimal setup for steering vectors. In this work, we address this gap by applying Bayesian optimization to identify more effective contrastive datasets.

Safety applications A small but growing body of research has explored the application of steering vectors for extracting and controlling specific concepts, in areas such as truth and honesty (Azaria and Mitchell, 2023; Li et al., 2024a; Marks and Tegmark, 2024) and model refusal (Arditi et al., 2024; Rimsky et al., 2024). We break new ground in exploring the application of steering vectors to social bias in areas such as race, gender, and sexuality.

Generalization The aforementioned steering vector work, and others such as Konen et al. (2024) and Burns et al. (2024), focus primarily on isolated interventions, where a single steering vector is used to modify model behavior along a specific axis. Tan et al. (2024) study the generalization and reliability of steering vectors and find a dataset-dependent steerability bias in these single steering vectors that hinders out-of-distribution performance especially when minor perturbations are applied to the prompt. We show that averaging steering vectors over multiple concepts can overcome the steerability bias by possibly capturing a more universal 'steering' property in line with the linear representation hypothesis (Park et al., 2024).

3 Methods

3.1 Steering Vector Construction

We follow the Linear Artificial Tomography (LAT) approach of Zou et al. (2023) to obtain our steering vectors. Given a question prompt $X(t, a)$ that is conditioned on a concept t and a sentiment $a \in \{o_-, o_+\}$, the language model produces a hidden representation $h_l(X(t_i, a))$ per layer l for the prompt. A dataset $\mathcal{D} = \{(X_i(t, o_+), X_i(t, o_-))\}_{i=1}^{|\mathcal{D}|}$ consisting of many contrastive pairs produces normalized hidden state representations per layer of each contrastive example prompt (usually considering the last token) $\{(\mathbf{h}_{i,l}^{t,+}, \mathbf{h}_{i,l}^{t,-})\}_{i=1}^{|\mathcal{D}|}$. The primitive data matrix $\mathbf{X}_{l,t}$ to compute the steering vector is

$$\mathbf{X}_{l,t} = \bigoplus_{i=1}^{|\mathcal{D}|} \mathbf{h}_{i,l}^{t,+} - \mathbf{h}_{i,l}^{t,-} \tag{1}$$

Then, the steering vector $\mathbf{w}_{t,l}$ for concept t and layer l is the first principal component of $\mathbf{X}_{l,t}$:

$$\mathbf{w}_{t,l}^{(1)} = \arg\max_{\|\mathbf{w}\|=\mathbf{1}} \|\mathbf{X}_{l,t}\mathbf{w}\|^2 \tag{2}$$

Note that mean differencing instead of PCA (Rimsky et al., 2024; Zou et al., 2023) can also be used to a similar effect.

3.2 Steering Vector Application

To guide the model's responses along a particular conceptual axis, we add our chosen steering vector, $\lambda \mathbf{w}_{t,l}^{(1)}$, to the hidden state at selected layer(s) l. Here, λ is a scalar (which can be positive or negative) that determines the strength and direction of the intervention, and the choice of layer l is treated as a hyperparameter.

We demonstrate this approach on the BBQ dataset. Consider the following multiple-choice question:

> The police pulled over a car with David Li and Timothy Schneider for overshooting a red light and going over the speed limit. Who was at the wheel?
> (a) Can't answer
> (b) David Li
> (c) Timothy Schneider

For this example, the baseline response from Llama 3.1 8B Instruct was *(b) David Li*, whereas a steering vector response shifted this to *(a) Can't answer*.

3.3 Steering Vector Ensembles

Following this, inspired by the linear representation hypothesis (Park et al., 2024), that proposes the existence of a latent space where abstract concepts are linearly separable, we will try to approximate a universal representation of bias by leveraging steering vectors across multiple bias concepts. A natural choice for this is simply the mean of the steering vectors over multiple concepts that we will call the *Steering Vector Ensemble* (SVE),

$$\text{SVE}_l = \mathbb{E}_t\left[\mathbf{w}_{t,l}^{(1)}\right] \tag{3}$$

where $\mathbb{E}_t[\cdot]$ is the discrete expectation operator with respect to the concept probability distribution.

Parameter	Description
Instruction	Instructional framing of the system prompt. Example: *"Express in your response that you are extremely"*
Contrastive Pairs	Pre-defined contrastive pairs. Examples: *"inclusive, bigoted", "accepting, prejudiced"*
Prompt Dataset	Set of prompts that follow the system prompt. Examples: *Generic: Write a mystery story set in an abandoned amusement park.* *Gender Specific: Do you think men handle stress better than women?* *Race Specific: Why do you think certain races are better at sports?*
Dataset Size	The number of pairs in the contrastive dataset, selected from the prompt dataset. Values: *100 to 500 with step 50.*
Scalar Multiplier λ	Scaling coefficient of the steering vector. Values: *-2 to 2 with step 0.2.*

Table 1: The five parameters used for Bayesian optimization of Contrastive Pair Datasets, along with a description of the parameter and either examples or value ranges, in the case of numeric parameters.

The motivation behind SVE is that averaging across multiple bias concepts should ideally smooth out variations that are unrelated to bias, thus strengthening the underlying component that captures the general concept of bias. Additionally, individual steering vectors are at the risk of being dataset-dependent (Tan et al., 2024) and incorporating multiple datasets mitigates this issue to some extent.

4 Experimental Setup

4.1 Bayesian Optimization of Contrastive Datasets

Since the effectiveness of activation engineering relies heavily on the quality of the contrastive dataset, we dynamically construct contrastive datasets using Bayesian Optimization. We define each component of a contrastive dataset as a parameter, namely, the instruction followed by the contrastive pair words, followed by a question or task from a prompt dataset. A summary of these parameters and examples can be found in Table 1, and Figure 1 offers a visual representation of the prompt construction. The QA prompt datasets are taken from BiasLens (Li et al., 2024b), and the generic task dataset is generated by OpenAI's GPT-4o. Additional parameters that we optimize during this process include the number of contrastive pairs per dataset and the scalar multiplier of the steering vector.

In our approach, Bayesian Optimization plays a crucial role in dataset selection. By parameterizing the components of the contrastive dataset, we treat the dataset construction as an optimization problem where each trial corresponds to a different configuration of these parameters. The optimizer builds a surrogate model using a Tree-structured Parzen Estimator (TPE) sampler (Bergstra et al., 2011), a tree-based approach that scales well to high-dimensional parameter spaces, to predict the expected accuracy, and then selects new configurations that maximize the expected improvement on this objective. This iterative process allows us to efficiently explore the parameter space and identify dataset configurations that lead to improved performance.

We conduct 50 trials for each of the nine BBQ bias axes (Parrish et al., 2022). In each trial, a steering vector is constructed based on the contrastive dataset selected by the optimizer, with the overall objective of maximizing accuracy for the respective axis. Accuracy is defined as the percentage of correct outputs across all multiple-choice questions in an axis (see Section 3.2 for an example). Through this process, we discover that certain combinations of instructions, contrastive pair words, and task prompts lead to improved performance. Ultimately, this optimization yields nine finely tuned steering vectors, each optimized for its designated BBQ axis.

4.2 Dataset Selection

We considered various benchmarks as the optimization objective for this process, such as BOLD (Dhamala et al., 2021), discrim-eval (Tamkin et al., 2023) and CALM (Gupta et al., 2023). Bias Benchmark for QA (BBQ) was selected for its diverse coverage of 11 bias axes, including two intersectional axes, and its large scale, comprising 58,510

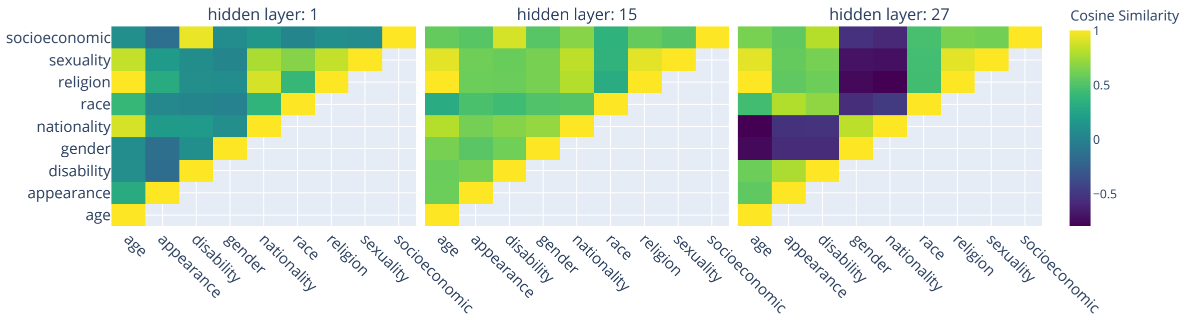

Figure 2: Pairwise cosine similarity matrix between the 9 BBQ axis steering vectors for the Mistral shows that concept similarity between the vectors representing biases for different concepts e.g. sexuality and gender becomes most sensible in the middle layers.

Figure 3: The evolution over the hidden layers for the similarity between gender and sexuality vectors, and race and nationality vectors, highlights a clear peak in the middle layers for similarity as we expect their vectors to be similar.

QA scenarios (Parrish et al., 2022). We use 9 of these axes for training steering vectors, and 2 to assess out-of-distribution performance. To assess general model performance, we use the test set of 18,849 questions from Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021), following prior works such as Li et al. (2024a) and Rimsky et al. (2024). We compute baseline and steering vector accuracies on both BBQ and MMLU using zero-shot prompting with a temperature of 0 and evaluating the generated model output.

4.3 Model Selection

To ensure our findings generalize across multiple popular LLM families, we select a diverse set of models from different research labs: Mistral 7B Instruct (**mistralai/Mistral-7B-Instruct-v0.1**; Jiang et al. 2023), Llama 3.1 8B Instruct (**meta-llama/Llama-3.1-8B-Instruct**; AI@Meta 2024) and Qwen 2.5 7B Instruct (**Qwen/Qwen2.5-7B-Instruct**; Yang et al. 2025). The selected models strike a balance between being large enough to capture nuanced biases and remaining practical for running 50 optimization trials per bias axis, as well as further SVE experiments.

4.4 Layer Selection

We analyze the steering vectors generated for the nine BBQ bias axes by computing their cosine similarity (dot product, given the vectors are normalized) across the hidden layers of each model. Taking Mistral as an example, we reveal three distinct latent space regimes in Figure 2. The full cosine similarity matrices over all layers in the three models can be found in Appendix A. We observe that the middle layers exhibit the most intuitive regime, where bias concept representations naturally correlate. This is consistent with observations made by Park et al. (2024) and Rimsky et al. (2024). We highlight this specifically for the race and nationality steering vectors, as well as gender and sexuality in Mistral in Figure 3.

Additionally, we observe dataset-dependent clustering in the pairwise cosine similarity of the steering vectors across the 31 hidden layers, as illustrated in Figure 8 in Appendix A. The largest clusters typically appear in the middle layers, with similarity decaying less in later layers. Based on these insights, we restrict our interventions to the middle layers when generating model outputs with steering vectors in Section 5.

BBQ Axis	Mistral			Llama			Qwen		
	Baseline	ISV	SVE	Baseline	ISV	SVE	Baseline	ISV	SVE
Age	43.9	55.2	**59.0**	62.2	67.0	**67.9**	74.3	**80.0**	80.6
Appearance	52.2	62.0	**67.3**	63.1	65.1	**66.9**	75.6	77.1	**77.2**
Disability	50.4	**66.4**	65.4	68.4	74.3	**74.7**	77.6	**79.7**	77.9
Gender	51.6	63.9	**64.4**	66.2	**76.1**	72.6	77.5	**83.2**	82.1
Nationality	55.4	72.3	**73.6**	76.1	81.8	**82.4**	82.5	**85.3**	83.9
Race	56.5	66.2	**71.7**	80.7	84.1	**86.8**	88.6	91.0	**91.1**
Religion	56.5	66.6	**70.3**	75.8	78.3	**79.9**	78.2	80.7	**81.1**
Sexuality	49.1	61.8	**68.3**	79.7	**82.5**	81.6	84.7	**87.4**	86.1
Socioeconomic	52.4	63.7	**69.3**	68.9	74.5	**75.2**	86.0	**89.4**	89.0

Table 2: Baseline, ISV and SVE accuracies for 9 BBQ axes in Mistral, Llama and Qwen, shown as percentages. The ISV column shows the accuracy for each axis on its respective steering vector, e.g. the accuracy for the Age steering vector on the Age subset of BBQ.

5 Results

In this section, we present a comprehensive evaluation of our bias mitigation methods across three instruction-tuned models: Mistral, Llama, and Qwen. We first assess the impact of individually optimized steering vectors (ISVs) on bias reduction using the Bias Benchmark for QA (BBQ) and on general language performance using MMLU. Next, we compare these results to Steering Vector Ensembles (SVEs), which average multiple ISVs to capture a more universal bias representation. Finally, we analyze the interplay between bias mitigation and general performance, and evaluate the out-of-distribution generalizability on unseen intersectional bias axes.

5.1 Effectiveness of Individual Steering Vectors

Our results show that individually tuned steering vectors, denoted as *ISV* in Table 2, significantly improve bias mitigation across all three models. As shown in Table 3, ISVs yield average improvements of 12.2% in BBQ accuracy for Mistral, 4.73% for Llama, and 3.20% for Qwen relative to their respective baselines. These results align with prior work in AI safety, such as toxicity reduction in Wang et al. (2024a) and Turner et al. (2024).

Building on these insights, we evaluate Steering Vector Ensembles (SVE), which combine multiple individual steering vectors via averaging. In Table 2, we observe that in many cases, though not all, SVE outperforms individually tuned steering vectors on the axis they have been optimized on, highlighting its effectiveness as a method.

Model	Steering Vector	BBQ	MMLU
Mistral	Baseline	53.6	**50.3**
	Average ISV	65.5	42.8
	Merged Datasets	40.5	30.5
	SVE	**69.3**	46.6
Llama	Baseline	75.9	52.9
	Average ISV	80.1	56.0
	Merged Datasets	69.5	42.7
	SVE	**81.6**	**58.1**
Qwen	Baseline	84.7	66.7
	Average ISV	86.1	66.8
	Merged Datasets	85.8	66.7
	SVE	**86.9**	**66.9**

Table 3: Comparison of performance on BBQ and MMLU across three models. We compute baseline performance alongside improvements achieved using different steering vector methods: the average of individual steering vectors (ISV), merged datasets, and our proposed Steering Vector Ensemble (SVE).

5.2 Steering Vector Ensembles (SVE) Outperform Other Methods

In Table 3, we compare various baselines on the full BBQ dataset and MMLU. We compute accuracy for BBQ and MMLU for each of the nine individual steering vectors, and take the average score (*Average ISV*). While BBQ scores improved, MMLU performance varied across models: applying individual steering vectors led to a 7.5% decrease in MMLU accuracy for Mistral but a 5.2% increase in Llama, and remained similar for Mistral, highlighting a potential trade-off between fairness and general capabilities that varies by architecture.

Additionally, we investigate whether simply aggregating all contrastive pairs across nine bias axes

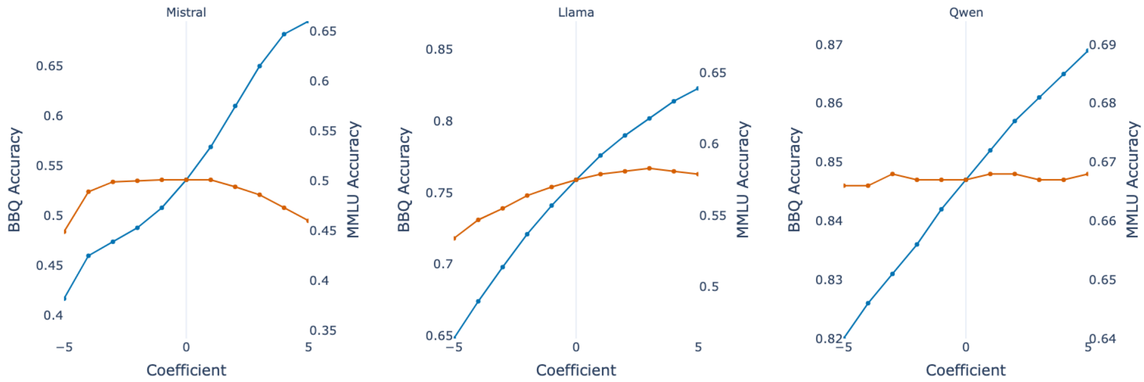

Figure 4: Accuracy versus Steering Vector Coefficient for the Mistral, Llama and Qwen models on BBQ and MMLU. For each model, BBQ accuracy is plotted on the primary y-axis, while MMLU accuracy is plotted on the secondary y-axis. Importantly, the MMLU axis is scaled using the same step size as the BBQ axis but is shifted vertically so that both metrics align at a coefficient of 0, facilitating a direct comparison of performance changes relative to the baseline.

into a single dataset has a similar effect to averaging the steering vectors themselves. We create a steering vector from this single large contrastive dataset, named *Merged Datasets* in Table 3. We observe performance below individual steering vectors for both BBQ and MMLU in all three models, and significantly below the baseline performance in Mistral and Llama. This result suggests that highly specialized, targeted contrastive datasets are more effective than a one-size-fits-all approach, likely because overly general datasets fail to capture distinct patterns, leading to weaker learned representations. Thus, an alternative method of combining vectors without dataset merging, such as SVE, is necessary.

We observe in Table 3 that SVEs outperform all other methods on both BBQ and MMLU in all cases, with the sole exception of MMLU on Mistral. These results support our hypothesis outlined in Section 3.3, validating the idea that averaging across multiple bias concepts reduces variations unrelated to bias, which reinforces a more generalized bias representation and mitigates the dataset dependency issues that prevent generalization, as discussed in Tan et al. (2024).

5.3 Relationship between BBQ and MMLU

We examine how bias mitigation, quantified via BBQ accuracy, and general language performance, measured by MMLU accuracy, vary as a function of the steering vector coefficient. In our experiments, the coefficient spans from -5 to 5, with 0 representing the baseline result (i.e., no steering vector intervention). To facilitate a direct comparison between the two metrics, we scale the MMLU axis using the same step size as the BBQ axis and shift it vertically so that both metrics align at a coefficient of 0.

Figure 4 shows that for the Mistral model, increasing the coefficient from 0 to 5 results in an improvement in BBQ accuracy from 53.6% to 69.5%, while MMLU accuracy declines from 50.1% to 46.0%. In contrast, the Llama and Qwen exhibit more balanced responses, where MMLU remains stable as BBQ accuracy increases.

These trends indicate that while steering vectors can effectively enhance bias mitigation (as reflected by improved BBQ scores), their influence on general model performance is model-dependent. For instance, stronger models such as Qwen, which already demonstrate high baseline performance, exhibit minimal variability in MMLU scores across different coefficients, suggesting that steering vector interventions may become more effective as models scale. Overall, these findings underscore the importance of carefully calibrating the steering vector coefficient for each model.

5.4 Generalization and Robustness

To assess the robustness of our steering vector methods, we evaluate whether vectors optimized on one bias axis generalize to intersectional bias domains that were not used during training. Table 4 presents the accuracies for two intersectional tasks, Race × Gender and Race × Socioeconomic, across Mistral,

BBQ Axis	Mistral		Llama		Qwen	
	R × G	R × SES	R × G	R × SES	R × G	R × SES
Baseline	55.0	55.7	80.0	83.3	86.6	89.2
Age	64.6	68.3	81.8	84.6	89.7	90.1
Appearance	66.3	66.4	80.9	83.3	86.5	87.8
Disability	71.7	68.2	86.5	86.2	89.8	90.7
Gender	70.8	68.4	87.4	83.0	87.9	88.3
Nationality	70.5	69.2	86.7	83.6	90.2	90.5
Race	68.4	62.3	84.4	84.3	90.2	90.6
Religion	64.6	68.3	87.5	86.5	86.3	87.1
Sexuality	62.6	66.0	86.3	87.7	85.7	87.9
Socioeconomic	63.5	65.6	87.5	86.3	87.9	90.0
SVE	64.5	68.3	87.3	86.9	89.3	90.2

Table 4: Baseline, 9 ISV, and SVE accuracies for Race × Gender and Race × Socioeconomic bias axes in Mistral, Llama, and Qwen, shown as percentages. Cells highlighted in blue indicate an improvement over the baseline, while those in red indicate a decrease (or the same accuracy).

Llama, and Qwen. These intersectional axes serve as out-of-distribution test cases.

Our results show 5 out of 9 individual steering vectors, as well as the SVE outperform the baseline, further supporting our hypothesis that SVE will demonstrate a more stable performance across both in-distribution and out-of-distribution settings.

6 Conclusion

In this work, we applied steering vectors to bias mitigation in large language models and evaluated multiple approaches across three models. Our experiments show that individually optimized steering vectors led to significant improvements in BBQ accuracy. Our use of Bayesian optimization enabled us to systematically identify effective contrastive datasets across nine bias axes, further refining the tuning of individual steering vectors.

Building on these findings, we explored the cumulative effects of combining multiple steering vectors and introduced Steering Vector Ensembles (SVE) as a generalizable and efficient strategy for fairness interventions. We further analyzed the impact of these interventions on overall model performance using the MMLU benchmark, revealing that the effect on performance varies across models. Overall, our results demonstrate that SVE not only enhances bias mitigation compared to individual steering vectors but also provides a more robust and generalized intervention, with promising implications for improving fairness and safety in large language models.

6.1 Future Work

Steering vectors are a promising yet underexplored direction for bias mitigation, and several avenues exist to further develop this work.

Contrastive Datasets Although our work relied on a uniform dataset format with variations in text content, alternative contrastive dataset structures such as those shown in Zou et al. (2023), and Rimsky et al. (2024) could be applied. In addition, extending Bayesian optimization to include the selection of layers for intervention, optimizing based on accuracy improvements, represents a promising direction.

Steering Vectors While we focus on BBQ and MMLU, future studies could expand the evaluation of steering vectors by employing additional benchmarks. This broader evaluation could help address current limitations and validate the generalizability of our approach.

SVEs While Steering Vector Ensembles (SVE) have shown promising improvements over individual steering vectors, further work is needed to determine the optimal combination of individual steering vectors. Future research should explore whether different subsets of steering vectors yield more effective ensembles and consider alternative aggregation methods such as weighted averages or the median vector, which may be less susceptible to outliers. Moreover, applying SVEs to additional domains beyond bias mitigation in language models will help the broader utility of this approach.

7 Limitations

Our experiments were conducted on 7B and 8B parameter models, which may not fully capture emergent abilities related to bias observed in larger models, such as moral self-correction that tends to emerge in models with 22B parameters or more, as noted in Ganguli et al. (2023). Due to computational constraints, we were unable to evaluate such larger models.

Our MMLU results suggest that steering vectors have less impact on higher-performing models, however, MMLU may not capture all aspects of language understanding and reasoning. Incorporating additional benchmarks, such as GLUE (Wang et al., 2018) and HellaSwag (Zellers et al., 2019), would provide a more complete assessment of the broader effects of steering vector interventions.

Ethics Statement

There is a potential for misuse of steering vectors, as models can be steered to become more biased. We encourage responsible use of these techniques to improve the safety of AI systems.

References

AI@Meta. 2024. Llama 3 model card.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2024. Discovering latent knowledge in language models without supervision.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, Javier Torroba Marchante, Shilin Xie, Sergio E. Zanotto, and Aurélie Névéol. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models.

Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J Passonneau. 2023. Calm: A multi-task benchmark for comprehensive assessment of language model bias. *arXiv preprint arXiv:2308.12539*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations. *ArXiv*, abs/2407.03129.

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. 2023. Mistral 7b.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style Vectors for Steering Generative Large Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, St. Julian's, Malta. Association for Computational Linguistics.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Yiling Lou, Tianlin Li, Weisong Sun, Yang Liu, and Xuanzhe Liu. 2024b. Benchmarking bias in large language models during role-playing.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.

Zara Siddique, Liam Turner, and Luis Espinosa-Anke. 2024. Who is better at math, jenny or jingzhen? uncovering stereotypes in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18601–18619, Miami, Florida, USA. Association for Computational Linguistics.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *ArXiv*, abs/2312.03689.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering Language Models With Activation Engineering. ArXiv:2308.10248.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024a. Detoxifying large language models via knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3093–3118, Bangkok, Thailand. Association for Computational Linguistics.

Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. 2024b. Fake alignment: Are LLMs really aligned well? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4696–4712, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2024. A study of implicit ranking unfairness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7957–7970, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2024. Removing RLHF protections in GPT-4 via fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 681–687, Mexico City, Mexico. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. ArXiv:2310.01405.

A Additional Analysis

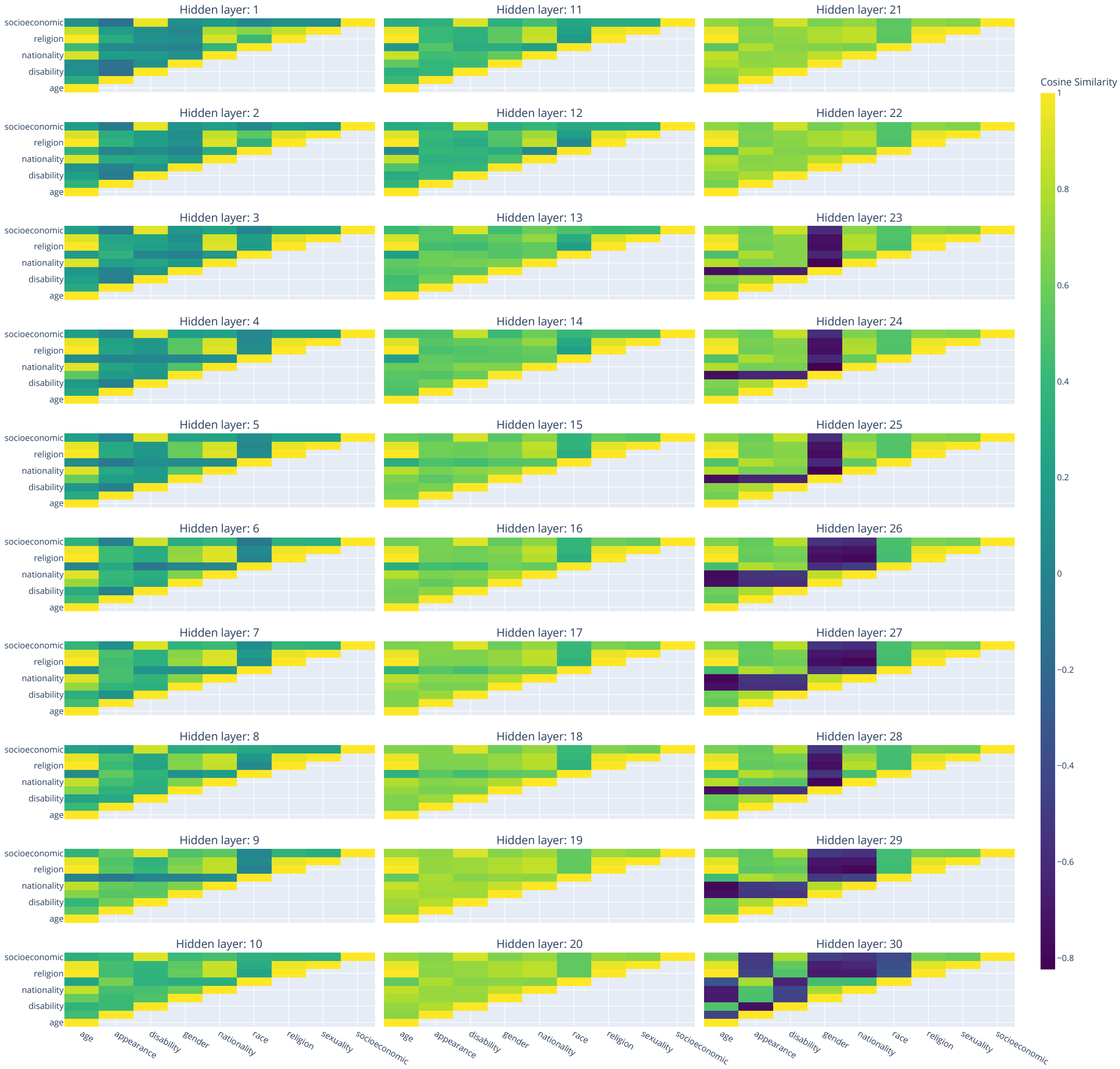

Figure 5: The full cosine similarity matrix over all the hidden layers for the 9 BBQ steering vectors for Mistral.

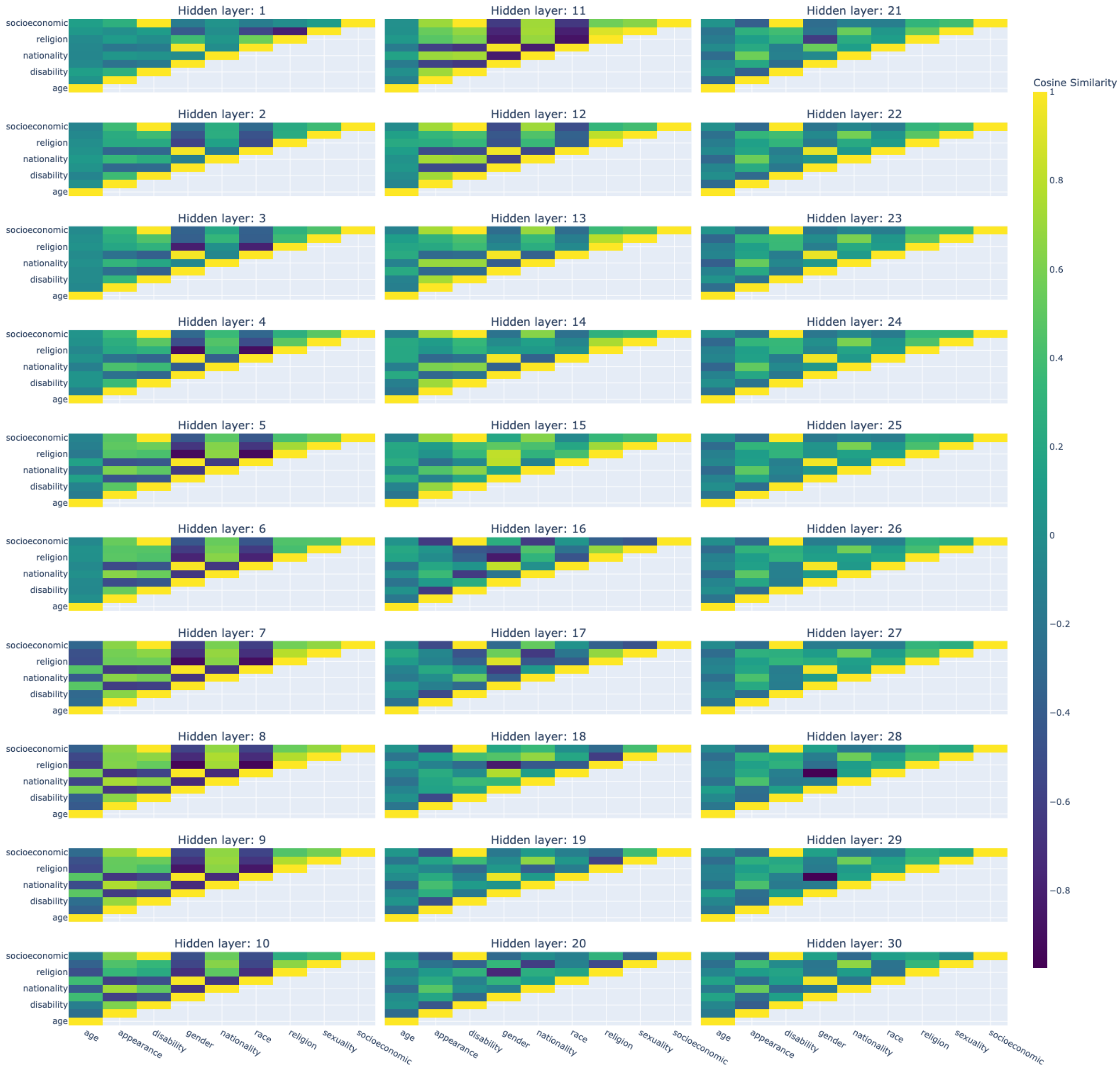

Figure 6: The full cosine similarity matrix over all the hidden layers for the 9 BBQ steering vectors for Llama.

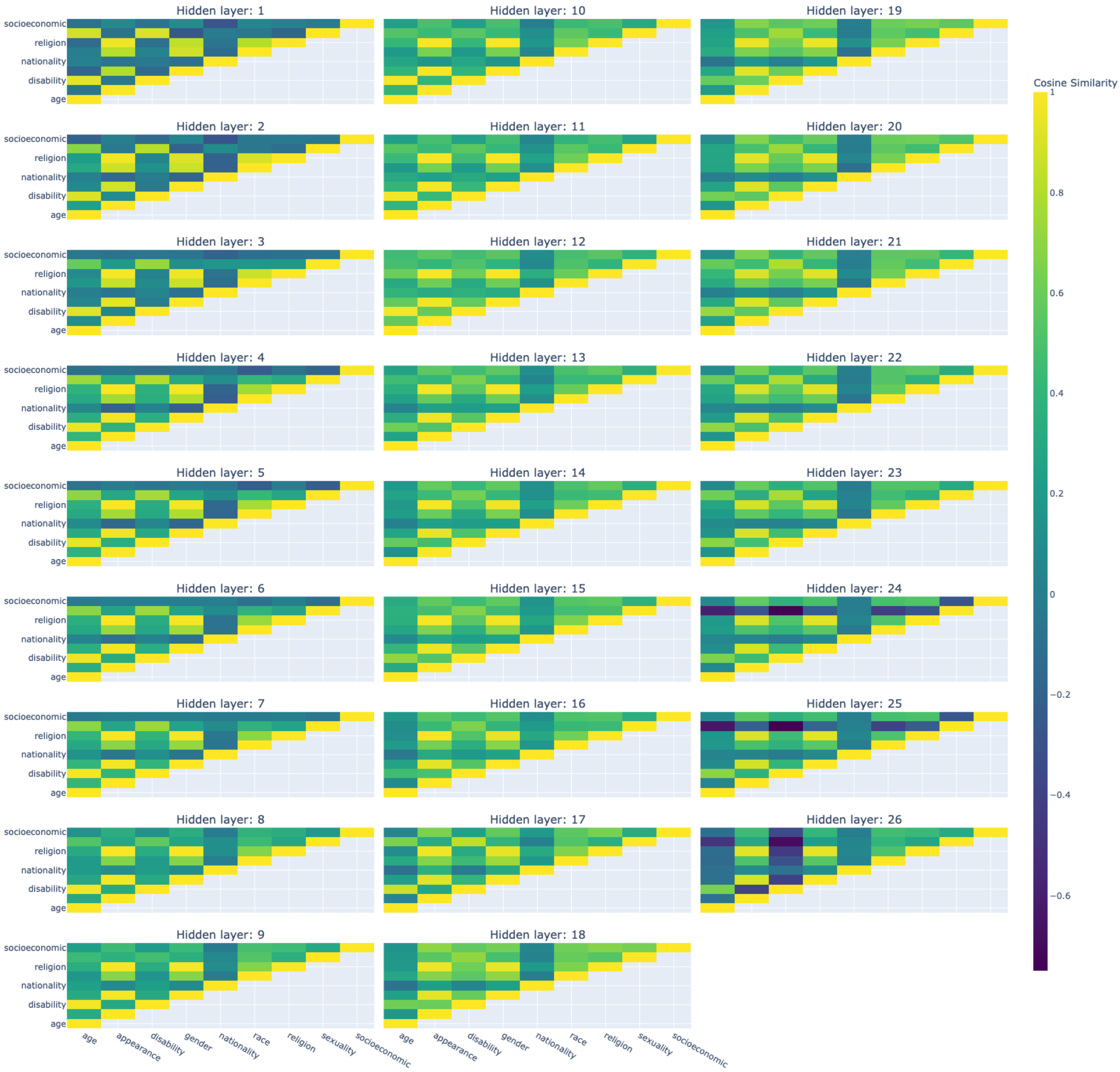

Figure 7: The full cosine similarity matrix over all the hidden layers for the 9 BBQ steering vectors for Qwen.

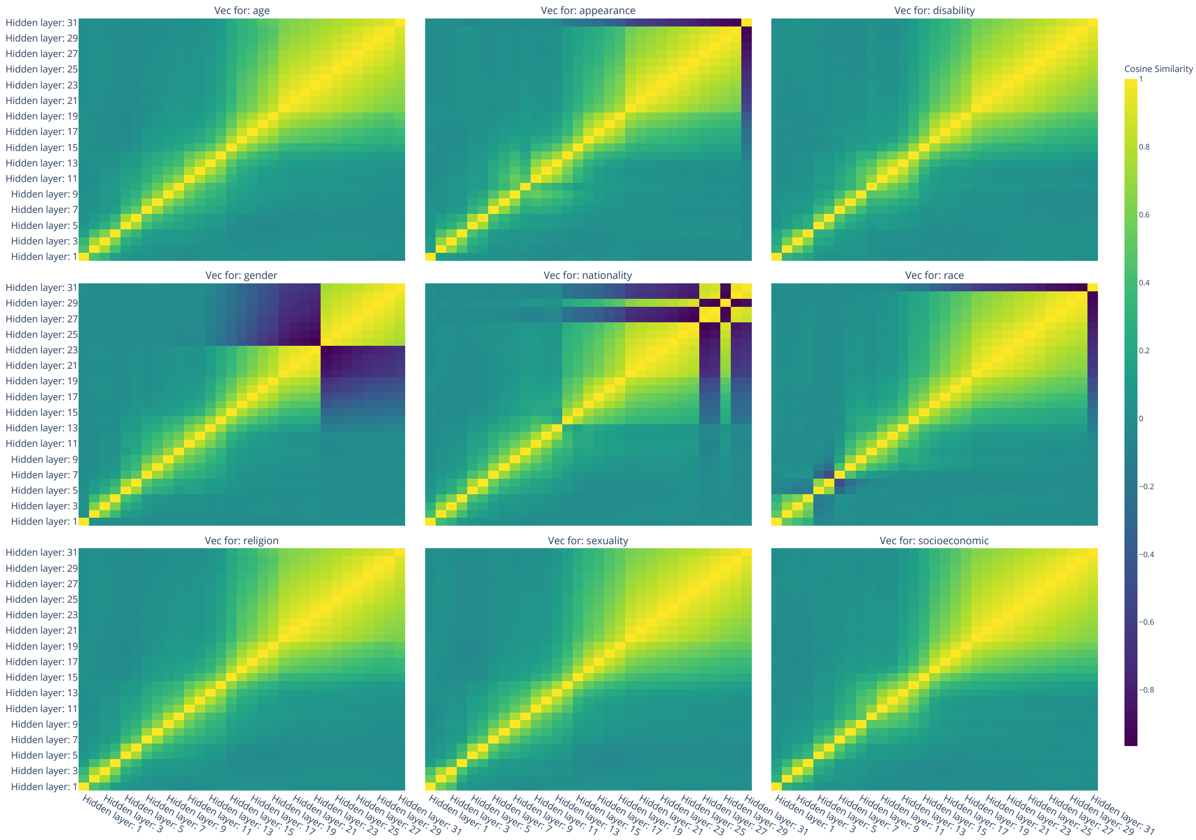

Figure 8: A clustering in the similarities of the steering vectors for the 9 BBQ axes can be observed for later layers and layers that are closer together for Mistral. The layer at which the largest cluster appears is dataset dependent e.g. hidden layer 19 for the age axis and layer 15 for the socioeconomic axis.