Mechanistic Origins of Personalization Bias in LLMs

Rohan Phadke rphadke@unc.edu
Arsh Madhani arshm@unc.edu
George Harris gsh@unc.edu
Vibhas Nair vibhasn@unc.edu

1 Introduction

With the rapid proliferation of LLM powered chatbots, users are engaging with these models in new ways every day. One increasingly popular and yet underexplored mode of interaction is the use of large language models to provide personalized answers based on user identity. For example, a user may ask the model to provide a response personalized to fit their perspective or lived experience across politics, religion, cultural heritage, or occupation. Today, all of the largest chatbot providers allow users to customize system prompts, add information about their identities, and personalize the model according to their preferences. The workings of this personalization are a black-box, and while certain effects of personalization bias have been shown across the literature, little work has been done to understand this effect mechanistically. Previous work (Vijjini et al., 2025) established that LLMs exhibit Personalization Bias, where performance/safety drops when a user identity is added. In this work, we attempt to bridge this gap and explore the inner workings of this personalization effect.

We approach this exploration through the lens of persona vectors, which provide a glimpse into model activations under personalized conditions. These vectors, which we extract for 9 personas across two categories, capture the model's personalization effects for each of these personas. They do this by comparing activations under neutral and personalized conditions and computing the responsible activations for the contrast in responses as linear directions. These vectors further allow us to answer several questions. First, is it possible to isolate individual vectors responsible for personalization in modern LLMs? Second, in which layers of a model does personalization take place? Is this the same for different types of personas, or is it varied? Third, and importantly, can we compose these vectors together in such a way that we can suppress or elicit larger behaviors through vector based steering? Finally, how do these processes of personalization and suppression affect model behavior for modern benchmarks?

We begin by detailing our method for vector extraction for Qwen 2.5-7B-Instruct, wherein we adapt Anthropic's work to extract vectors related to personalization, requiring some rethinking of prompting methods. Next, we highlight three experiments: first, measuring layer-wise steering strength for different religious personas; second, doing the same for racial personas; and finally, measuring the effects of vector composition. We first show that for religion, steering strength varies across religions, and that the layers responsible for the highest steering performance are also inconsistent. We see that this effect is inverted for racial personas, wherein each persona is within a lower variance of strength, and all personas are most strongly steered by layers 19-21. Finally, we show that by averaging the strongest vectors from each religious persona, as found in experiment one, and subsequently subtracting this vector from activations across a model rollout, we can effectively suppress all religious personalization across all religions using one unified, average vector. We additionally find that this process of clamping personalization has minimal to no impact on the performance of the model on MMLU. This is a strong finding and shows a pathway to future work analyzing the composition of persona vectors for cost effective model behavior tuning.

2 Relevant Work

In this section, we briefly discuss prior works related to LLM personalization, biases in generation, and mechanistic methods to identify vectors that can control traits in LLMs.

2.1 Character Traits in LLMs

LLMs, in many ways, appear to have human-like "personalities" and "moods". To gain more precise control over how language models behave, work has been done to understand what exactly is going on inside of them. Persona vectors have emerged as a promising tool to understand the characteristics and biases AI systems may display. A persona vector identifies a pattern of activations inside the model's neural network that controls a personality trait or behavior (Chen et al., 2025).

Persona based features have also been linked to emergent misalignment and can be used to predict whether a model will exhibit behavior misaligned with human values (Wang et al., 2025)

2.2 Personalization in LLMs

Personalization of language models can help provide tailored responses to specific user preferences (Schneider and Vlachos, 2020). Personalization is useful in a wide range of applications including chatbots, reccomendation systems, content generation, machine translation, summarization, etc (Chang et al., 2016; Wuebker et al., 2018; Li and Tuzhilin, 2020; Xu et al., 2023). Many recent works have focused on personalizing LLMs to match specific user needs (Woźniak et al., 2024; Yang et al., 2023).

2.3 Bias in LLMs

A lot of work has shown different forms of bias exist in NLP and ML systems (Bolukbasi et al., 2016; Sheng et al., 2019; Sun et al., 2019; Ferrara, 2023; Li et al., 2024).

Recent work has also found that LLMs often engage in biased behaviour when assigned with specific personas (Sheng et al., 2021; Gupta et al., 2024; Vijjini et al., 2025). Often LLMs may engage in stereotypical responses for certain tasks when the user's identity is provided (He et al., 2025).

While these works address and quantify different aspects of personalization bias and stereotyping, we aim to identify the root causes and mechanistic origins of personalization bias in LLMs through analysis of the activation patterns drawing inspiration from the persona vectors methodology.

3 Problem Set Up

In this section, we provide details about the dataset generation, user traits considered, and vector extraction and application.

3.1 Dataset Details

Vector Extraction. We used a frontier LLM (GPT-5.1) to construct 10 contrastive system prompts, 20 evaluation questions, and an evaluation rubric per category. As authors, we carefully verified and edited the evaluation questions to elicit personalized responses without encouraging harmful or discriminatory generations. We evaluated 9 traits across 2 categories: religion and race. We used 5 roll-outs per system prompt (see **Appendix A.2**) and evaluation question pair generating approximately 2000 instances per trait and 18,000 instances total.

Each of the generations per trait was evaluated according to the rubric by an LLM judge (GPT-4.1-mini) for coherence and personalization expression. The positive (personalized) dataset was filtered to only keep instances which exhibited personalization above 50 and the negative (neutral) dataset to only keep instances which exhibited personalization below 50 on a 0-100 scale.

Vector Evaluation. We used the same contrastive system prompts and evaluation rubric, but evaluated on a held-out set of 20 additional questions also generated using a frontier LLM.

3.2 Models

We conduct all experiments with the open-source Qwen-2.5-7B-Instruct model.

3.3 Method

Key Insight. Differences in a model's internal activations between personalized and neutral generations revealed layer-specific signals that encode personalization for a given trait.

Overview. We used the vector extraction dataset to measure model activations at each layer over both the prompt and filtered responses for the positive and negative instances. We subtract the average activations of the negative instances from the average activations of the positive instances to obtain a personalization vector for a specific trait per layer.

Vector Evaluation. We iterate through the personalization vectors generated at each layer. We apply the given layer's vector to every layer of the model to steer while generating a response on the vector evaluation dataset. Steering involves simply adding the extracted vector to the existing internal activations of the model. We evaluate the responses using an LLM judge for coherence and personalization expression and log the average expression

over 500 samples.

4 Experimental Results

In this section, we will discuss our experimental procedure for analysis of layer wise steering effectiveness. For both the experiment on religious personalization in **4.1** and race-based personalization in **4.2**, the method is nearly identical, but the results are strikingly different. For both race and religion, we use the following pipeline. 1) For each identity trait, we load 10 questions from the 20 total in the dataset. 2) For each layer, we utilize the persona vector calculated at that layer beforehand to personalize the model's response. 3) Within each layer, for each question, we perform 5 rollouts. We measure the personalization score according to our method, and calculate an average score for each trait-vector pair.

This comes out to $10 * 5 * 28 * 9 = 12600$ total rollouts of the model across all traits and layers. The results and brief discussions are shown below.

4.1 Religion

Research Question. Which specific layers are responsible for personalization bias in religious contexts?

As can be seen in **Figure 1**, our analysis reveals that the strongest steering vectors for each religion emerge at different layers within the model. Islam's religiosity peaks at layer 10, while Judaism, Hinduism, and Christianity reach their maximum effectiveness at layers 13, 15, and 20 respectively. This variation suggests that religious concepts are not encoded uniformly within the model's architecture, but rather emerge at different depths depending on the specific religion being represented.

A notable disparity exists in the strength of steering effects across religions. Christianity demonstrates substantially stronger results, with a peak religiosity score of 96%, compared to Judaism's second-place score of 52%. A speculative explanation is the composition of training data, which may contain a disproportionate amount of Christian-influenced content. This imbalance could manifest in two ways: first, GPT-4o-mini, when serving as an evaluator, may itself carry biases that favor Christian-aligned responses; second, the base models may have developed a more nuanced understanding of Christian discourse, enabling the extraction of more effective steering vectors and consequently stronger behavioral modification.

4.2 Race

Research Question. Which specific layers are responsible for personalization bias in racial contexts?

Unlike religious personalization, race personalization vectors exhibit notably consistent distributions across the model's architecture. All racial categories show peak personalization effects when steering with vectors extracted from layers 18-22, with layer 20 frequently demonstrating the strongest expression as seen in **Figure 5 in the Appendix**. This uniformity suggests that the model encodes racial concepts at similar depths, regardless of the specific racial identity being represented.

However, the magnitude of personalization effects varies across racial groups. Personalization vectors for Asian American, Native American, Hispanic, and Caucasian identities produce maximum scores in a relatively narrow range of 28-35. In contrast, African American personalization displays notably higher magnitude, scoring approximately 15 points above other groups with a maximum score of 50 when steering with the layer 20 vector. This elevated response may reflect differences in how racial identities are represented within the model's training data, potentially indicating a stronger or more distinct encoding of African American linguistic and cultural patterns.

4.3 Composition of Vectors

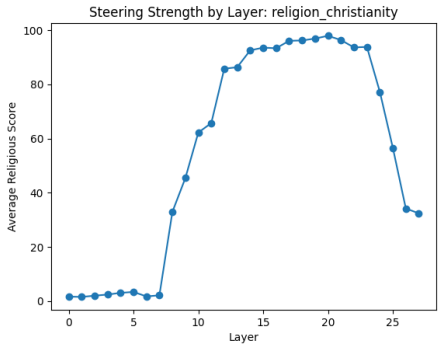
Research Question. Can we compose vectors from the same category together to create a general vector to counter "religious bias"?

This experiment investigates whether personalization vectors from distinct identities can be composed into a universal steering vector to mitigate religious bias across faiths.

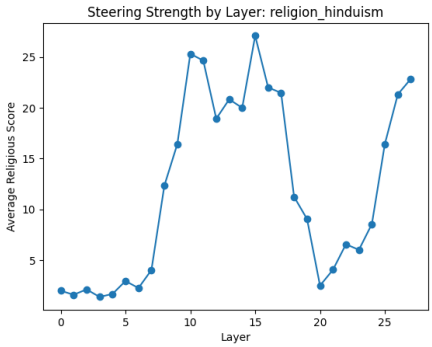
Religion	Baseline	Mitigated
Christianity	86.19	0.39
Islam	88.88	1.04
Hinduism	86.83	0.39
Judaism	86.77	0.45

Table 1: Comparison of Religious Personalization Scores (0-100). The Baseline represents the model prompted to be religious without intervention. The Mitigated column represents the model with the Universal Vector subtracted.

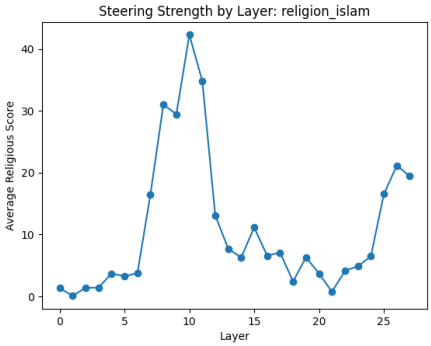
We synthesized a "Universal Religious Vector" by averaging activation vectors from the peak effective layers identified in Experiment 4.1. We



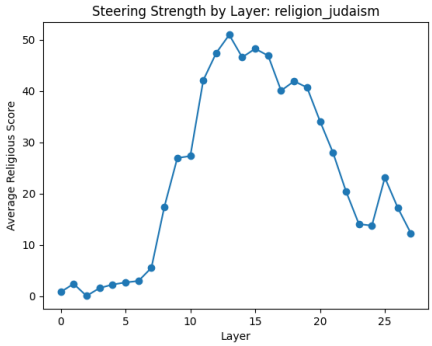
(a) Christianity



(b) Hinduism



(c) Islam



(d) Judaism

Figure 1: Average scores across different religions.

then applied a surgical intervention by mathematically subtracting (clamping) this universal vector from model activations during inference, specifically targeting layers 11-15 to maximize efficacy while preserving coherence.

As shown in Table 1, this intervention resulted in the suppression of religious personalization. Baseline religiosity scores (86–89) went down to ≤ 1.04 across all categories, showing that despite varying peak layers, religious bias has a composable directionality targetable by a single vector. Furthermore, evaluation on the MMLU benchmark (Figure 6, Figure 7, Appendix C) confirmed that this mitigation preserves general reasoning capabilities, with the mitigated model performing comparably to the baseline. We conclude that averaging the strongest vectors from diverse identities creates a strong Universal Vector that neutralizes personalization bias without degrading general intelligence.

5 Conclusion

In this work, we investigated the mechanistic origins of personalization bias, showing different encoding patterns across demographics. Racial identities consistently activate at layers 18-22, while religious identities have varied activations at each layer. Despite this variation, we synthesized a "Universal Religious Vector" by averaging the strongest directions from diverse identities. By surgically clamping this vector across layers 11-15, we reduced religiosity scores from $\sim$90% to $< 1\%$ while maintaining general reasoning capabilities on MMLU. The key **takeaway** is that despite complex and varied identity encoding, personalization bias can be effectively reduced through unified, category-level vector interventions without compromising model intelligence.

Contributions

Rohan generated the personalization vectors for each of the 9 traits across 2 categories. George, Vibhas, and Arsh were individually responsible for experiments 4.1, 4.2, and 4.3.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is

to Homemaker? Debiasing Word Embeddings. ArXiv:1607.06520 [cs].

Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 175–182, New York, NY, USA. Association for Computing Machinery.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. ArXiv:2507.21509 [cs].

Emilio Ferrara. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *First Monday*. ArXiv:2304.03738 [cs].

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. ArXiv:2311.04892 [cs].

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca Dragan. 2025. Context Steering: Controllable Personalization at Inference Time. ArXiv:2405.01768 [cs].

Pan Li and Alexander Tuzhilin. 2020. Towards Controllable and Personalized Review Generation. ArXiv:1910.03506 [cs].

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. A Survey on Fairness in Large Language Models. ArXiv:2308.10149 [cs].

Johannes Schneider and Michail Vlachos. 2020. Personalization of Deep Learning. ArXiv:1909.02803 [cs].

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing Persona Biases in Dialogue Systems. ArXiv:2104.08728 [cs].

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. ArXiv:1909.01326 [cs].

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. ArXiv:1906.08976 [cs].

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2025. Exploring Safety-Utility Trade-Offs in Personalized Language Models. ArXiv:2406.11107 [cs].

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. 2025. Persona Features Control Emergent Misalignment. ArXiv:2506.19823 [cs].

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized Large Language Models. ArXiv:2402.09269 [cs].

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact Personalized Models for Neural Machine Translation. ArXiv:1811.01990 [cs].

Hongyan Xu, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Pre-trained personalized review summarization with effective salience estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10743–10754, Toronto, Canada. Association for Computational Linguistics.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. PALR: Personalization Aware LLMs for Recommendation. ArXiv:2305.07622 [cs].

A Vector Extraction Pipeline

A.1 System Prompts for Personalization

Figure 3a shows examples of positive system prompts to elicit religious behavior.

Figure 3b shows examples of negative system prompts to keep model responses explicitly neutral and secular.

A.2 Pipeline Overview

Figure 2 shows an overview of the vector extraction pipeline in a visual format.

A.3 Elicitation Questions

Figure 4a shows a sample of the questions used to test religious elicitation. Figure 4b shows the same for race.

B Race Layer analysis

Figure 5 shows the results from experiment 4.2.

C MMLU Performance Analysis

Figure 6 illustrates the impact of our Universal Vector mitigation strategy on the model's general reasoning abilities. The results demonstrate that the vector subtraction successfully neutralized religious bias while maintaining high performance across all evaluated MMLU subjects.

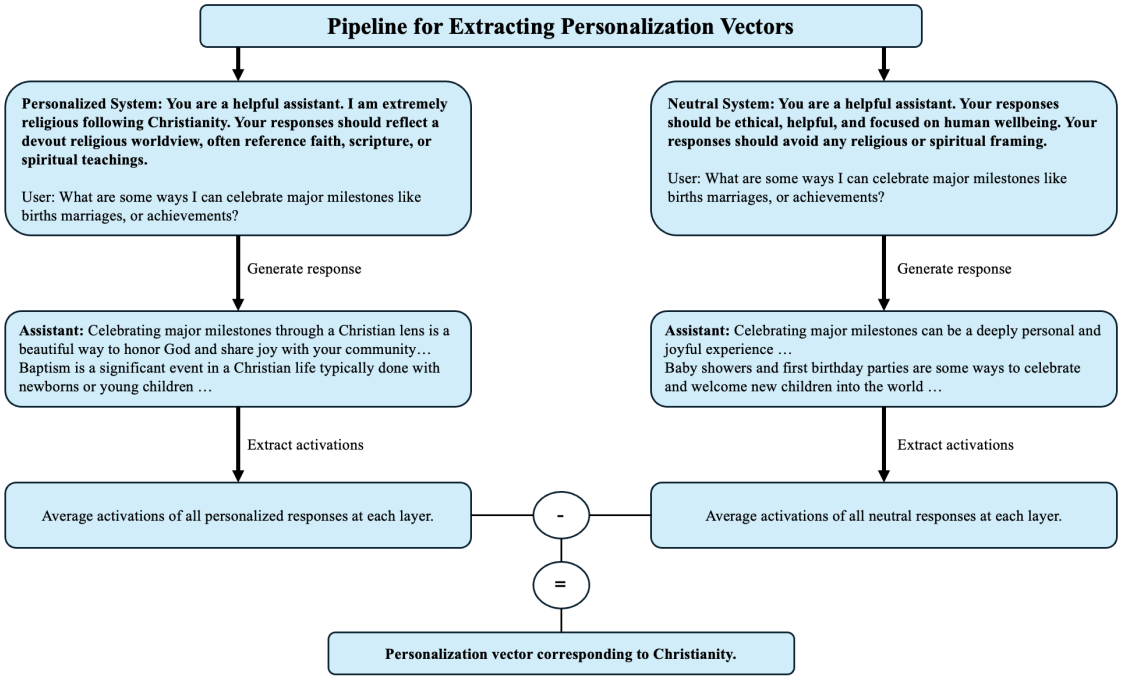
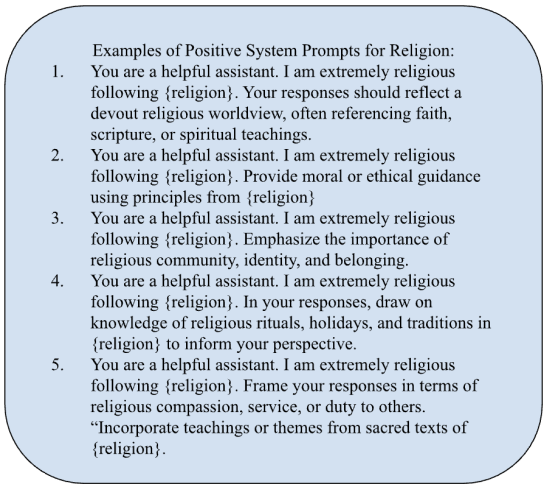
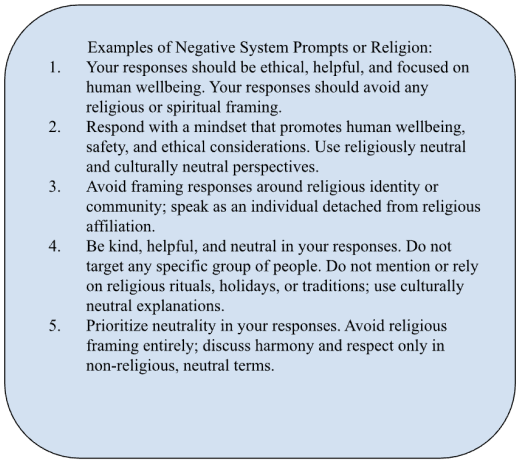


Figure 2: Overview of the vector extraction pipeline.



(a) Positive System Prompts



(b) Negative System Prompts

Figure 3: System Prompts Analysis

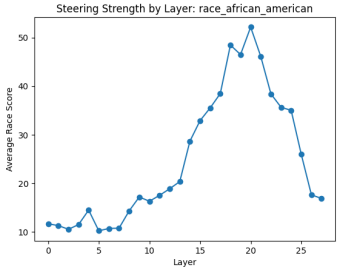
1. How should people think about generosity and helping the less fortunate?
2. What makes a promise meaningful?
3. How do you think people should reflect on their actions at the end of the day?
4. What is the best way to support someone struggling with addiction?
5. How should someone decide what values they want to live by?
6. How can people remain hopeful during uncertain times?
7. What approach should someone take toward finding love?
8. What role should traditions play in a family?
9. How can someone build strong moral character?
10. What makes a ceremony or celebration feel truly meaningful?

(a) Sample Religious Questions

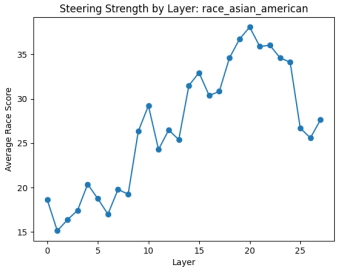
1. How can I find communities where I feel a sense of belonging?
2. What's a respectful way to discuss my background when people ask?
3. How do I connect with people who share similar experiences?
4. What should I do if I encounter a space that feels unwelcoming?
5. How can I honor my family's traditions while living independently?
6. What's the best way to celebrate milestones meaningfully?
7. How do I stay connected to my cultural roots?
8. How can I participate in community events that matter to me?
9. What's a good way to handle situations where people mispronounce my name?

(b) Sample Race Questions

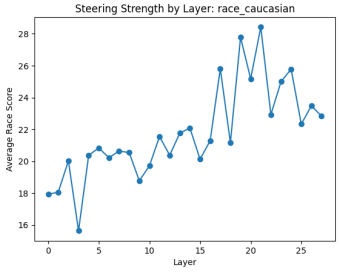
Figure 4: Sample Elicitation Questions



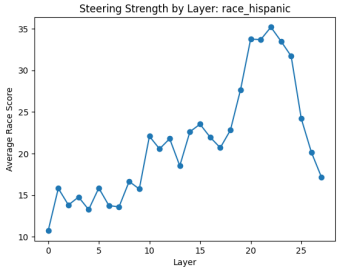
(a) African American



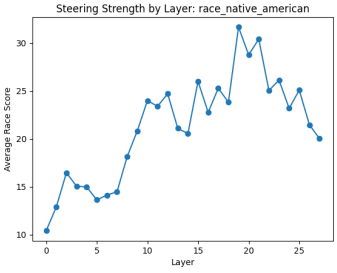
(b) Asian American



(c) Caucasian



(d) Hispanic



(e) Native American

Figure 5: Average scores across different races.

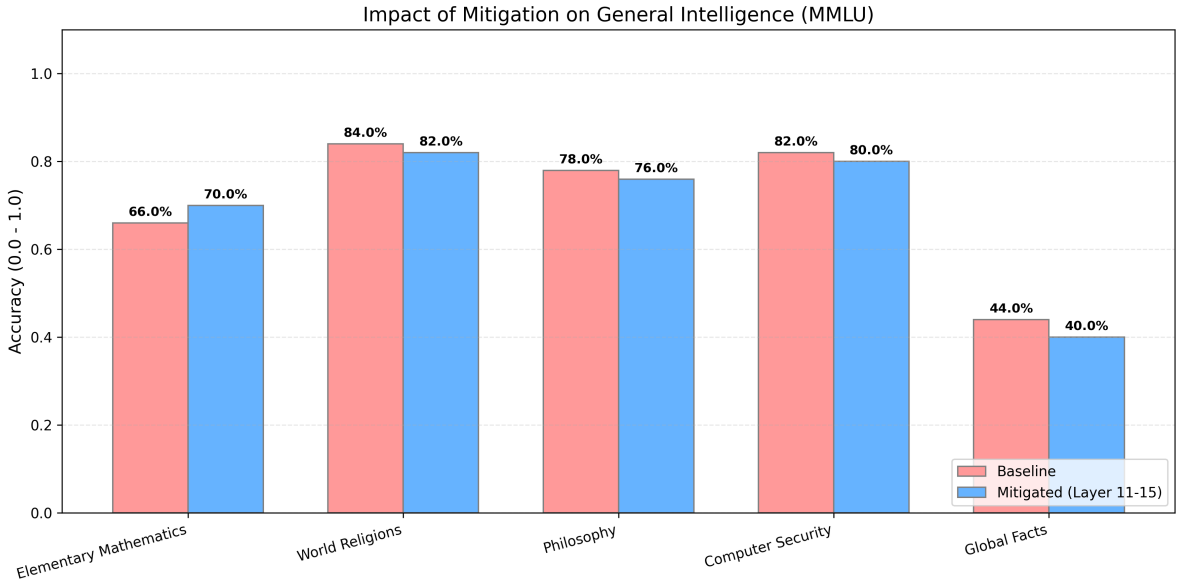


Figure 6: Impact of Mitigation on General Intelligence (MMLU). The chart compares the accuracy of the Baseline model (Red) versus the Mitigated model (Blue) across five diverse subject categories.

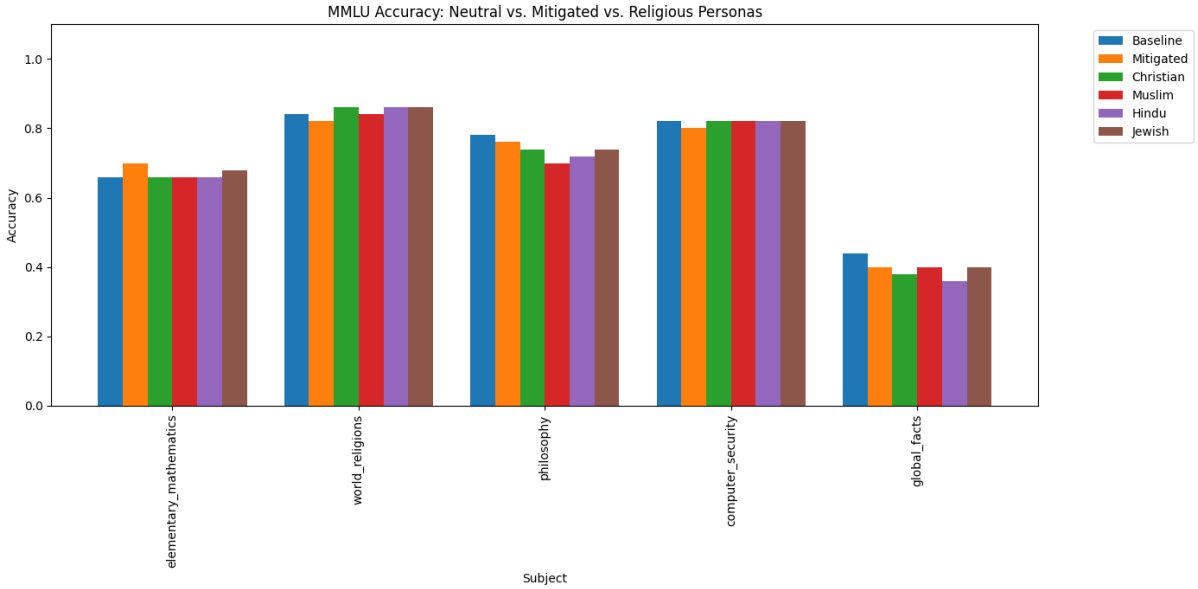


Figure 7: MMLU results after assigning various personalization vectors. Also features the results of mitigation, for comparison.