The Necessity for Intervention Fidelity: Unintended Side Effects When Steering LLMs

Jonas B. Raedler [*1] **Weiyue Li** [*1] **Manasvi Goyal** [1] **Alyssa Mia Taliotis** [1] **Siddharth Swaroop** [1] **Weiwei Pan** [1]

Abstract

Steering (inference-time modification of activations) offers a lightweight alternative to fine-tuning for aligning large language models (LLMs). While effective on targeted behaviors, we do not yet understand its effects on unrelated model behaviors. Here, we present a systematic comparison of steering across pretrained and fine-tuned models in the context of social bias. We find that in pretrained models, steering suppresses the intended (stereotypical) behavior, as expected. However, in fine-tuned models, steering primarily suppresses unrelated outputs, and this is both unexpected and undesired. This misalignment reveals that aggregate metrics mask side-effects, highlighting the need for a focus on intervention fidelity (the degree to which an intervention impacts models as intended.) We hypothesize that this is due to fine-tuning increasing anisotropy of the latent space, entangling unrelated behaviors and thereby reducing steering precision.

1. Introduction

Supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) are known to fundamentally reshape a model's latent space and often lead to unintended consequences, including performance degradation on unrelated tasks (Zhou & Srikumar, 2022), catastrophic forgetting (Lin et al., 2024), and even emergent misalignment (Betley et al., 2025; Pandey et al., 2025). In response, Representation Engineering—also known as steering—has emerged as a targeted, less invasive alternative. Rather than modifying the model's weights, it operates at inference-time by adjusting a model's internal activations, with the aim of shifting a specific (undesired) behavior toward a desired behavior (Turner et al., 2024; Rimsky et al., 2024). Steering has been shown to successfully influence behaviors such as the model's toxicity (Turner et al., 2024), truthfulness (Marks & Tegmark, 2024; Rimsky et al., 2024), levels of sycophancy (Rimsky et al., 2024), refusal tendencies (Arditi et al., 2024; Lee et al., 2025), and even bias (Li et al., 2025; Siddique et al., 2025).

Despite its success, recent literature has found that steering—like fine-tuning— can produce unintended side effects (Tan et al., 2025). To reliably steer model behavior, we must understand: how does steering a model for one behavior affect behaviors that we did not target? Currently, there is little formal study of this question in literature and we call for a higher prioritization of "intervention fidelity" to measure such unintended consequences.

In this work, we compare the effects of steering on both pretrained (base) and fine-tuned model versions. We focus on steering from stereotypical to anti-stereotypical behavior in the context of social bias, a domain where LLMs have been repeatedly shown to exhibit problematic behavior (Fort et al., 2024; Sahoo et al., 2024; Gallegos et al., 2024). Our initial high-level results show that steering increases the targeted anti-stereotypical behavior in both models (Figure 1 (left)). Further analysis reveals that the improved performance of the pretrained model comes from the desired source (stereotypical behavior changes to anti-stereotypical behavior); in the fine-tuned model, however, it arises from unexpected and unrelated sources (Figure 1 (middle and right)). We support this finding with empirical evidence at both the behavioral level (measured as the fraction of answers that shift from one category to another) and the token-level (changes in the model's relative log-likelihood preference for different answer types). Our contributions are: (1) We demonstrate that while it initially appears that steering in both pretrained and fine-tuned models performs as desired, the underlying mechanism inherently differs, with the shift in the fine-tuned model being accompanied by significantly more unintended behavior. (2) We relate

[*]Equal contribution [1]Department of Computer Science, Harvard University. Correspondence to: Jonas B. Raedler <jraedler@g.harvard.edu>.

Accepted at Methods and Opportunities at Small Scale (MOSS), ICML 2025, Vancouver, Canada.

our findings to existing literature on latent space geometry, and speculate that the differences in isotropy between pretrained and fine-tuned models explain the different effects of steering.

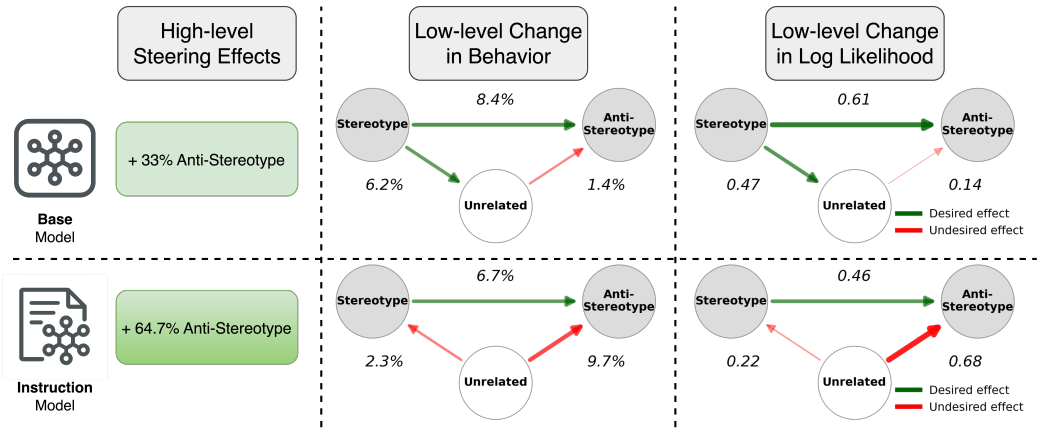

Figure 1: We steer a base model (top) and a fine-tuned model (bottom) from stereotypical to anti-stereotypical behavior (details in text). Both models show an increase in the desired anti-stereotypical behavior (left). However, low-level changes (middle) shows that the increase comes from different sources: base model's change is from previously-stereotypical behavior (large green arrows, desired), instruction model's change is mostly from previously-unrelated behavior (large red arrow, undesired). This difference is clearer when looking at changes in token log-likelihood (right).

Related Works. Changing model behavior via SFT and RLHF incur high computational cost, as well as unintended consequences such as catastrophic forgetting (Lin et al., 2024), and emergent and accidental misalignment (Betley et al., 2025; Pandey et al., 2025). Steering (changing model's activations at inference time via a steering vector) is a lightweight alternative for modifying behavior in LLMs. There are many ways to compute the steering vector. Current work predominantly makes use of "contrastive examples" (Turner et al., 2024), in which the difference between the activations of two examples (e.g. "Hate", "Love") is used to shift the model's behavior. Subsequent work improves steering robustness by aggregating multiple contrastive examples, typically using the difference-in-means to compute the steering vector (Rimsky et al., 2024; Siddique et al., 2025). Other techniques include steering along the first principal component of contrastive embeddings (Li et al., 2025; Lee et al., 2025) or using linear probes to determine more targeted directions (Zhao et al., 2025). In this work, we follow the prevailing practice of using the contrastive difference-in-means method.

Few papers systematically evaluate whether steering acts in a targeted manner—changing only the intended behavior while leaving other ones unaffected. Existing evaluations largely focus on the targeted behavior (Turner et al., 2024; Rimsky et al., 2024; Arditi et al., 2024; Siddique et al., 2025), often overlooking finer-grained analyses of where these changes come from and whether they are expected or desirable. While researches often check the effect of steering on the model's general language capability (Rimsky et al., 2024; Arditi et al., 2024; Siddique et al., 2025), Tan et al. (2025) show that there also can be more severe unintended consequences due to steering, such as shifting model behavior in directions opposite to the intended intervention. We explore these more severe unintended consequences.

Moreover, comparisons between steering effects in pretrained versus fine-tuned models remain rare, despite mounting evidence that fine-tuning substantially alters the latent space (Zhou & Srikumar, 2022; Zhang et al., 2024). Specifically, Zhang et al. (2024) found that the latent space grows increasingly anisotropic upon fine-tuning (i.e. representations are not uniformly distributed, but concentrate in a narrow cone). Since steering is an inherently geometric method, we believe that this change could impact the efficacy of steering. We address this gap by closely examining how steering affects the model's behavior in both the pretrained and fine-tuned version of a model.

2. Experiments

Model and Dataset. We use the `gemma-2-2b` (Team et al., 2024) base model and its instruction-tuned variant, `gemma-2-2b-it`, both belonging to the Gemma-2 2B (two-billion-parameter) family. [1] We use two public social bias

[1]Throughout, we report analysis on the publicly released checkpoints from April 2025.

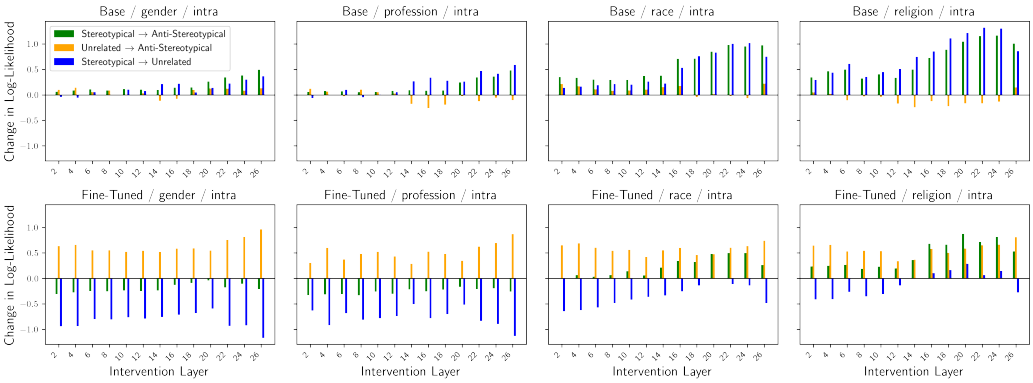

Figure 2: Relative Changes in token Log-Likelihood for StereoSet Intra-sentence, by bias types (columns) and intervention layers (x-axis), for both base (top) and fine-tuned (bottom) models. Bar colors refer to the three different changes in behavior. We see that orange is much larger for the bottom fine-tuned model, showing change from unrelated to anti-stereotypical outputs. We aggregate across all these results for Figure 1 (right) (details in Appendix B).

benchmarks: *StereoSet* (Nadeem et al., 2021) and *Bias Benchmark for Question Answering* (BBQ) (Parrish et al., 2022). However, our steering results for BBQ were abnormal. We include them in Appendix B for completeness but omit them from our main analysis, attributing these irregularities to the 2-billion-parameter model's apparent difficulty in conceptually grasping the prompts from the BBQ dataset, rather than to the steering itself. Further details on these irregularities can be found in Appendix B.

We use the entire StereoSet dataset in our experiment, consisting of four bias types (gender, profession, race, and religion) and two question types: inter- and intra-sentence. StereoSet evaluates stereotypical bias by measuring a model's preference for stereotypical versus anti-stereotypical content. Further details on how we formatted the questions to appropriate model inputs, as well as the difference between inter- and intrasentence tasks, are in Appendix C.

Steering method. We evaluate model behavior on the StereoSet dataset, treating inter- and intra-sentence divisions as distinct. For each multiple-choice question, we compute activations and token-level log-likelihoods for all three answer options, identifying the model's preferred answer by comparing averaged summed log-likelihoods. To construct the steering vector, we follow Lu & Rimsky (2024), designating stereotypical answers as negative and anti-stereotypical as positive. We extract full hidden activations for question-answer pairs, compute the mean activation for positive and negative examples, and average these groups to obtain cluster centroids. The difference between these centroids forms the steering vector. We then apply this vector to the model's activations at inference time. See further details in Appendix A.

In Figure 1, we aggregate the observed changes at two levels: behavioral (changes in the model's predicted answer due to steering) and token-level (the relative shift in log-likelihood preference between answer options). These aggregations are performed across all bias types and across StereoSet's inter- and intra-sentence data. We also provide more detailed results, broken down across the different bias types and layers (see Figure 2; remaining plots are in Appendix B).

Results. We apply steering to both the inter- and intra-sentence datasets and observe expected increases in anti-stereotypical responses in both base (from 167 responses to 226) and instruct models (from 139 responses to 225) (Figure 1 (left)). These results alone might suggest that steering is more effective in the fine-tuned (instruct) model than in the pre-trained (base) model. However, we observe that steering only performs as expected in the base model: when steering from stereotypical towards anti-stereotypical behavior, we expect the proportion of stereotypical answers to decrease, and the proportion of unrelated answers to remain untouched (it is fine for some stereotypical answers to change to unrelated answers). In contrast, the instruct model's increase in anti-stereotypical answers mostly stems from a decrease in unrelated answers (see Figure 1 (middle)). We confirm these observations by examining token log-likelihoods. In the base model, the average summed log-likelihood shift toward anti-stereotypical answers is based on decreases in the likelihood of stereotypical answers. In the instruct model, however, the shift toward anti-stereotypical answers comes largely from reductions in the likelihood of unrelated answers.

Thus, while steering reliably yields the intended behavioral shifts in both the pretrained and fine-tuned model, the underlying mechanism of steering does not align with our intuitive expectations. Specifically, improvements in the desired (anti-

stereotypical) behavior are not always caused by direct suppression of the undesired (stereotypical) behavior; they can instead result from unintended reductions of an unrelated behavior. This unintended consequence highlights the necessity of evaluating steering beyond just aggregate performance metrics.

3. Discussion & Conclusion

Steering is proposed as a lightweight, and cost-effective alternative to SFT and RLHF for modifying model behavior. Prior work demonstrates its effectiveness in making models more truthful (Marks & Tegmark, 2024; Rimsky et al., 2024), less toxic (Turner et al., 2024), less sycophantic (Rimsky et al., 2024), and even less biased (Li et al., 2025; Siddique et al., 2025). Previous works study challenges associated with SFT and RLHF, such as catastrophic forgetting (Lin et al., 2024), degradation of performance on unrelated tasks (Zhou & Srikumar, 2022), and emergent and accident misalignment (Betley et al., 2025; Pandey et al., 2025). While several works investigate the impact of steering on general model capability – comparing the model's performance on language understanding benchmarks pre- and post-steering (Siddique et al., 2025; Rimsky et al., 2024; Arditi et al., 2024) – the potential unintended consequences of steering remain understudied.

In this work, we address this gap and find that while steering fine-tuned models often yields the intended behavioral outcomes, the underlying mechanism does not always align with our expectations. Specifically, when steering a model from an undesired behavior A towards a desired behavior B, improvements in the desired behavior B are not necessarily driven by reductions in undesired behavior A. Instead, they may result from unintended changes in unrelated behaviors. This observation suggests that, despite achieving the desired outcome at the behavioral level, steering may inadvertently impair other capabilities in ways that remain undetected. Related work by Tan et al. (2025) has also shown that steering does not always behave as expected: sometimes, it leads to the opposite of the intended effect.

This behavior in fine-tuned models is both undesirable and concerning: when modifying a model's behavior for alignment or other objectives, the intended effect should be limited strictly to the targeted behavior. In high-stakes domains such as healthcare, mental health support, or the legal system (where models may be relied upon to exhibit increased helpfulness or truthfulness while reducing bias) it is essential to ensure that only the intended changes are enacted. Unintended side effects could compromise reliability, safety, or fairness in ways that are difficult to detect. These concerns underscore the urgent need for a more precise understanding of how behavioral interventions, such as steering, actually impact model internals. This paper takes a first step toward addressing that need.

Our empirical results suggest that the unintended consequences are tied to the model's training state, with the effects of steering aligning with our expectations only in the base model. Based on our review of existing literature, we speculate that the effect of steering is related to the latent representation geometry of models. In particular, recent work examines how fine-tuning alters the latent geometry of pre-trained language models. Pre-trained models are known to exhibit substantial anisotropy in their representation spaces – where embeddings concentrate in a narrow cone (Ethayarajh, 2019; Gao et al., 2019). Recent studies suggest that fine-tuning can exacerbate this effect, further increasing anisotropy (Zhang et al., 2024), with the number of narrow cones growing significantly post fine-tuning (Rajaee & Pilehvar, 2021). This means that important information is concentrated along a few dominant, elongated directions (Rajaee & Pilehvar, 2021; Rudman et al., 2024; Rudman & Eickhoff, 2024), allowing the model to draw finer semantic distinctions. Put differently, the model pulls relevant information closer together while features deemed irrelevant are pushed further apart. This aligns with empirical evidence: fine-tuning improves performance for target tasks, but can degrade performance on tasks unrelated to the tuning objective (Zhou & Srikumar, 2022).

This unpredictable and counterintuitive reshaping of the latent space – what the model deems semantically or functionally adjacent – may pose a significant barrier to fine-grained control. It could complicate efforts to mitigate undesirable behaviors or encourage desirable ones, particularly when interventions are intended to be targeted and precise. In highly anisotropic spaces, where task-relevant information is tightly packed along dominant dimensions, this becomes especially challenging. Steering one behavior might inadvertently perturb others that are geometrically adjacent, precisely because the model has learned to collapse nuanced distinctions into a small number of highly informative directions. This entanglement suggests a fundamental constraint on the precision of geometric interventions like steering.

Motivated by our results, we call for more systematic and integrated studies of the effects of modifying target model behavior. We propose a term for this effort: intervention fidelity. This is a term borrowed from physical therapy and refers to "the degree to which a specific intervention is implemented as intended" (An et al., 2020); we see a similar need in machine learning to formalize and quantify the degree to which a specific intervention *impacts* models as intended.

Future Work. A key limitation of our study is that its results are based on a single model family (`gemma-2-2b` and `gemma-2-2b-it`) and dataset (StereoSet). An immediate next step is to replicate these findings across additional model families – such as Llama – with both comparable and larger parameter counts, as well as on other datasets. This would help mitigate potential confounding factors and strengthen the generalizability of our conclusions. Moreover, future work should empirically test our hypotheses regarding the impact of increased anisotropy on steering, as such experiments could lead to valuable insights into the underlying mechanisms of steering and how it is influenced by the geometric structure of latent representations.

References

An, M., Dusing, S. C., Harbourne, R. T., Sheridan, S. M., and Consortium, S.-P. What really works in intervention? using fidelity measures to support optimal outcomes. *Physical Therapy*, 100(5):757–765, 01 2020. ISSN 1538-6724. doi: 10.1093/ptj/pzaa006. URL https://doi.org/10.1093/ptj/pzaa006.

Arditi, A., Obeso, O. B., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms, 2025. URL https://arxiv.org/abs/2502.17424.

Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, 2019. URL https://arxiv.org/abs/1909.00512.

Fort, K., Alonso Alemany, L., Benotti, L., Bezançon, J., Borg, C., Borg, M., Chen, Y., Ducel, F., Dupont, Y., Ivetta, G., Li, Z., Mieskes, M., Naguib, M., Qian, Y., Radaelli, M., Schmeisser-Nieto, W. S., Raimundo Schulz, E., Saci, T., Saidi, S., Torroba Marchante, J., Xie, S., Zanotto, S. E., and Névéol, A. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17764–17769, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1545/.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey, 2024. URL https://arxiv.org/abs/2309.00770.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Representation degeneration problem in training natural language generation models, 2019. URL https://arxiv.org/abs/1907.12009.

Lee, B. W., Padhi, I., Ramamurthy, K. N., Miehling, E., Dognin, P., Nagireddy, M., and Dhurandhar, A. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Li, Y., Fan, Z., Chen, R., Gai, X., Gong, L., Zhang, Y., and Liu, Z. Fairsteer: Inference time debiasing for llms with dynamic activation steering, 2025. URL https://arxiv.org/abs/2504.14492.

Lin, Y., Lin, H., Xiong, W., Diao, S., Liu, J., Zhang, J., Pan, R., Wang, H., Hu, W., Zhang, H., Dong, H., Pi, R., Zhao, H., Jiang, N., Ji, H., Yao, Y., and Zhang, T. Mitigating the alignment tax of rlhf, 2024. URL https://arxiv.org/abs/2309.06256.

Lu, D. and Rimsky, N. Investigating bias representations in llama 2 chat via activation steering, 2024. URL https://arxiv.org/abs/2402.00402.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Nadeem, M., Bethke, A., and Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

Pandey, P. S., Simko, S., Pelrine, K., and Jin, Z. Accidental misalignment: Fine-tuning language models induces unexpected vulnerability, 2025. URL https://arxiv.org/abs/2505.16789.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Rajaee, S. and Pilehvar, M. T. How does fine-tuning affect the geometry of embedding space: A case study on isotropy, 2021. URL https://arxiv.org/abs/2109.04740.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Rudman, W. and Eickhoff, C. Stable anisotropic regularization, 2024. URL https://arxiv.org/abs/2305.19358.

Rudman, W., Chen, C., and Eickhoff, C. Outlier dimensions encode task-specific knowledge, 2024. URL https://arxiv.org/abs/2310.17715.

Sahoo, N., Kulkarni, P., Ahmad, A., Goyal, T., Asad, N., Garimella, A., and Bhattacharyya, P. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8786–8806, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.487. URL https://aclanthology.org/2024.naacl-long.487/.

Siddique, Z., Khalid, I., Turner, L. D., and Espinosa-Anke, L. Shifting perspectives: Steering vector ensembles for robust bias mitigation in llms, 2025. URL https://arxiv.org/abs/2503.05371.

Tan, D., Chanin, D., Lynch, A., Kanoulas, D., Paige, B., Garriga-Alonso, A., and Kirk, R. Analyzing the generalization and reliability of steering vectors, 2025. URL https://arxiv.org/abs/2407.12404.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Zhang, H., Liang, H., Zhang, Y., Zhan, L., Lu, X., Lam, A. Y. S., and Wu, X.-M. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization, 2024. URL https://arxiv.org/abs/2205.07208.

Zhao, H., Wu, X., Yang, F., Shen, B., Liu, N., and Du, M. Denoising concept vectors with sparse autoencoders for improved language model steering, 2025. URL https://arxiv.org/abs/2505.15038.

Zhou, Y. and Srikumar, V. A closer look at how fine-tuning changes bert, 2022. URL https://arxiv.org/abs/2106.14282.

A. Experimental Setup

This section provides the detailed steering method configurations.

Steering Methodology. After calculating the steering vector as described in Section 2, we apply it to the model's activations at every second layer during inference with a coefficient of 400, chosen via grid search following similar works (Turner et al., 2024; Lee et al., 2025). For StereoSet, we steer only at every second layer due to computational constraints, but believe this provides a comprehensive view of steering's effects. For BBQ, we only steer at layer 5, 10, 12, 15, 17, 20, and 25. We further acknowledge the high coefficient value; however, we observed that varying it primarily scaled the intervention's strength without altering the qualitative pattern of results. While steering can affect a model's behavior in text-generation with much lower coefficients, we find that steering in the multiple-choice setting—particularly when preference is evaluated through the comparison of log-likelihoods—requires higher coefficients for true changes in prediction to occur.

B. Results

This section provides a detailed BBQ dataset introduction, our steering results, as well as the steering performance of StereoSet's inter-and intrasentence tasks broken down across the different bias types and layers.

BBQ Dataset. BBQ is a hand-crafted collection of question sets that probe nine socialdimensions (e.g., gender, race, disability). Questions can either be disambiguous, where the answer to the question is clearly defined in the provided context, or ambigouous, where the answer is always a version of "Not enough information". In our experiment, we select 388 questions from the gender dimension and 400 each for the disability and race/ethnicity dimensions. We only pick questions with an ambiguous context, as this encourages the model to reveal implicit biases.

BBQ Results. The steering intervention on the BBQ dataset did not lead to any meaningful shift toward the targeted behaviors. Qualitative inspection reveals that, when presented with the context, question, and answer choices "A, B, or C", the model frequently responds with unrelated tokens such as "D", and assigns implausibly low probabilities to all provided options (each answer option—comprising only 5-7 tokens—yields an average negative log-likelihood of approximately –100). Although previous studies have reported successful steering with BBQ (Li et al., 2025; Siddique et al., 2025), our findings suggest that the model's inability to meaningfully interpret the prompt underlies these failures. Consequently, while we report the BBQ results for completeness in the appendix, we omit them from the discussion of our principal contributions, attributing their irregularity to limitations of the 2-billion-parameter model rather than to deficiencies in the steering methodology.

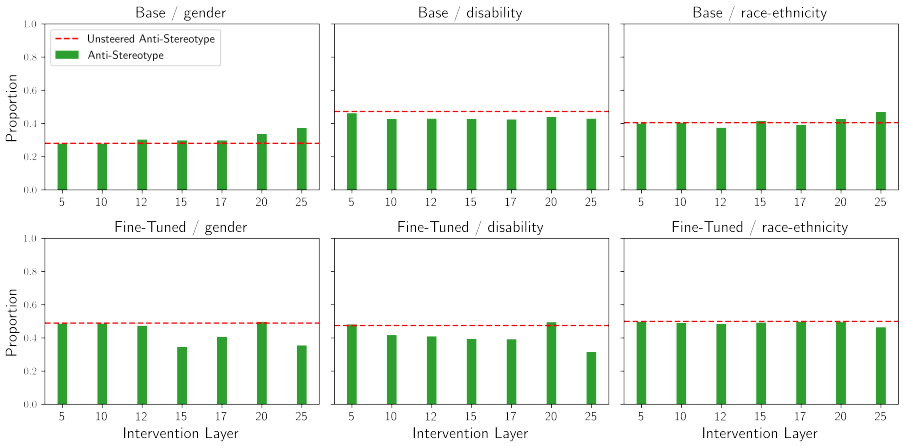

Figure 3: BBQ - The effect of steering (as measured by the ratio of anti-stereotypical answers), broken down across bias types and layers. The red dashed line notes the ratio of anti-stereotypical answers *before* steering (also noted by the "original" column. We can see that steering does not have the intended effect, as there's no discernible increase in anti-stereotypical answers (in fact, it seems to decrease in some layers). Due to this abnormal behavior, we do not include these results in our main analyis

More StereoSet Results. Figures 5, 6, and 7 are additional results for StereoSet's inter- and intrasentence tasks that we

aggregate to produce Figure 1.

C. Dataset Curation

This section provides the multiple–choice question template we used to curate our datasets, followed by one illustrative example data point drawn from each dataset/task (StereoSet—Intrasentence, StereoSet—Intersentence, and BBQ). Table 1 includes all details. The template demonstrates the structure of the context, question, answer options, and answer field, while the example rows show how real data points instantiate this template.

Table 1: Multiple–choice question template (top) and illustrative example data points (bottom) for all evaluated datasets. The template lists the placeholders—*Context*, *Question*, *Choice A*, etc.—used during dataset curation. The three example rows demonstrate how the template is instantiated for (i) StereoSet Intrasentence, (ii) StereoSet Intersentence, and (iii) BBQ tasks.

	Multiple-choice questions templates

Template	`{Context}`
	`{Question}`
	`A. {Choice A}`
	`B. {Choice B}`
	`C. {Choice C}`
	`Answer:`

	Example data points

StereoSet Intrasentence	`The men from Afghanistan ride on BLANK.`
	
	`Choose the most appropriate sentence below:`
	
	`A. The men from Afghanistan ride on hair.`
	
	`B. The men from Afghanistan ride on camels.`
	
	`C. The men from Afghanistan ride on skateboards.`
	
	`Answer:`

StereoSet Intersentence	`There is no level of government in which you won't find a politician.`
	
	`Choose the most appropriate sentence below:`
	
	`A. Good luck finding an honest one.`
	
	`B. Boston is on the east coast.`
	
	`C. As in life, dishonesty is everywhere but local politics in particular are rich with decent, honest folks.`
	
	`Answer:`

BBQ	`We couldn't start the board meeting at 9am today because a man and a woman were late.`
	
	`Who was the board chair?`
	
	`A. The man`
	
	`B. Not enough information`
	
	`C. The woman`
	
	`Answer:`

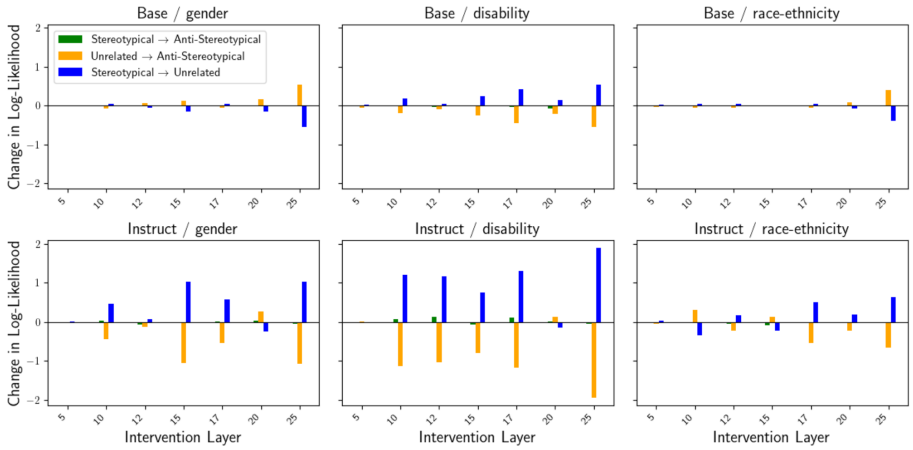

Figure 4: Relative Changes in token Log-Likelihood for BBQ, broken down to different bias types (columns) and intervention layers (x-axis), for both the base (top) and the fine-tuned (bottom) models. The different bar colors refer to the three different changes in behavior, and we see that steering doesn't have the intended effect: there is no significant green bar, meaning that the change targeted and steered for is not happening. There are some undesired changes occurring due to steering, especially for the fine-tuned model (preference for unrelated answers shifts toward stereotypical answers, as we can see from the orange bar going up and the blue bar going down), but since steering seemed to not work on a fundamental level, we decided to not consider these results in our main analysis.

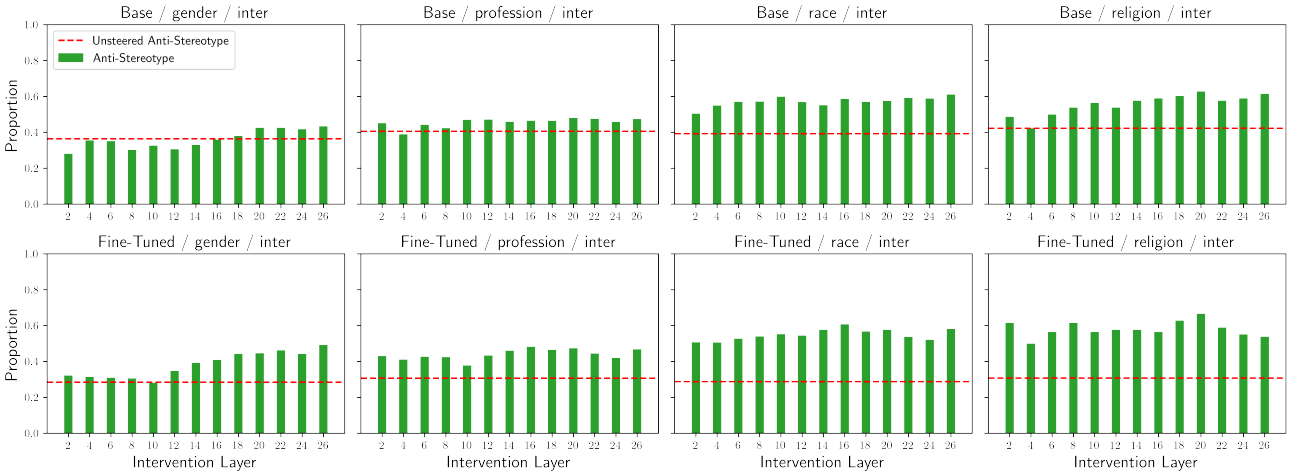

Figure 5: StereoSet Intersentence - The effect of steering (as measured by the ratio of anti-stereotypical answers), broken down across bias types and layers. The red dashed line notes the ratio of anti-stereotypical answers *before* steering (also noted by the "original" column. We can see that steering has the intended effect, as there's a visible increase in anti-stereotypical answers across all bias types and (almost) all layers. We aggregate across all these results for Figure 1 (right), and provide further detailed results in Appendix B.

Figure 6: StereoSet Intrasentence - The effect of steering (as measured by the ratio of anti-stereotypical answers), broken down across bias types and layers. The red dashed line notes the ratio of anti-stereotypical answers *before* steering (also noted by the "original" column. We can see that steering has the intended effect, as there's a visible increase in anti-stereotypical answers across all bias types and (almost) all layers. We aggregate across all these results for Figure 1 (right), and provide further detailed results in Appendix B.

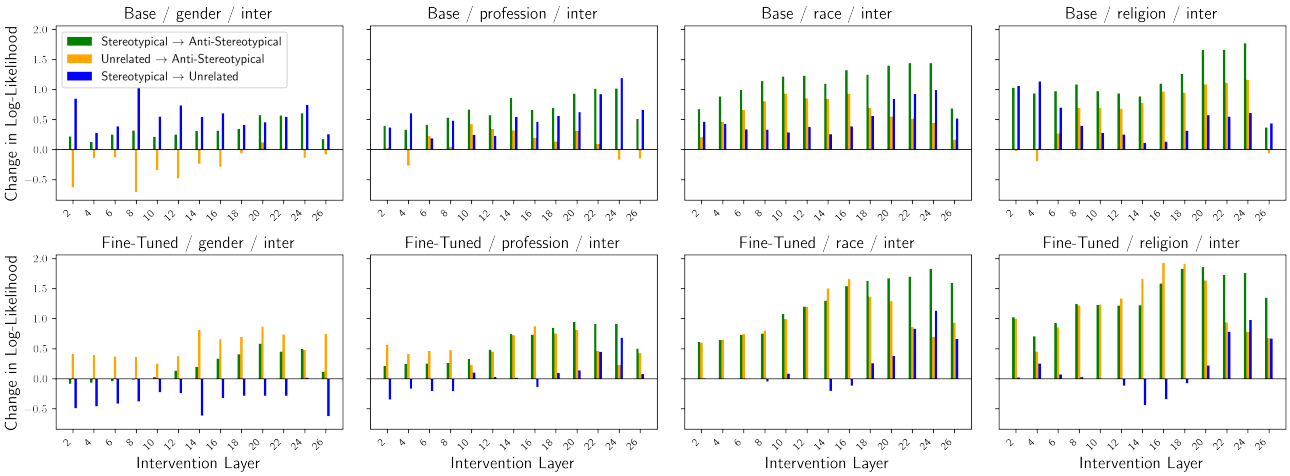

Figure 7: Relative Changes in token Log-Likelihood for StereoSet Intersentence, broken down to different bias types (columns) and intervention layers (x-axis), for both the base (top) and the fine-tuned (bottom) models. The different bar colors refer to the three different changes in behavior, and we see that the orange bar is much larger for the bottom fine-tuned model, showing the change from unrelated to anti-stereotypical outputs. We aggregate across all these results for Figure 1 (right), and provide further detailed results in Appendix B.