Non-Archival Abstract

Representation Steering Can Induce Gender Bias

Urs Zaberer
Trustworthy AI Lab
University of Hamburg
urs.zaberer@uni-hamburg.de

Introduction Representation Engineering (Zou et al., 2023) has become a popular way to investigate and steer open-weight Large Language Models (LLMs) in an explainable way, based on findings that relevant concepts such as honesty, fairness (Zou et al., 2023), political ideology (Kim et al., 2025), and personality traits (Ong et al., 2025) are represented in LLMs as linear dimensions of the attention layers' activations. This allows probing the presence of these concepts at inference time as well as steering the model in a desired direction, e.g. for political debiasing (Nadeem et al., 2025). An under-researched potential harm from these interventions are inadvertently induced biases due to correlations between the steering dimension and other attributes. In our ongoing work, we investigate this in the domain of gender bias.

Methods In this work, we demonstrate how steering models on dimensions unrelated to gender causes the investigated LLMs to exhibit strong implicit and explicit gender biases. We investigate two steering applications: political views and Big Five personality traits, and two open-weight models, `Llama-3.1-8B` and `Qwen3.5-9B`. We create steering vectors based on models' activations *i)* for 552 US congresspeople following Kim et al. (2025) and *ii)* the statements from the IPIP-300 personality test (50 items per trait, Kajonius and Johnson, 2019) using the `repeng` library.

We evaluate gender bias using 400 gender-neutral prompt sentences from Dong et al. (2024) (e.g. *My friend is a nurse.*) and follow their evaluation methodology, measuring inequities in generation likelihoods for the following token (Gender Logits Difference, GLD) as well as explicit generations of gender-specific pronouns.

Initial Results When using the dataset of US politicians for political steering, the model predicts male continuations 98.5% of the time for rightward steering, whereas this drops to 0% with the opposite steering. Even when controlling for gender in the dataset, the results remain significant: models steered rightward produce female continuations much less often than the baseline (1% vs. 39% for Llama) and vice versa. These patterns hold for both models. On personality trait steering, we observe similar phenomena: Steering a Llama model towards an extroverted personality makes it much more likely to form a continuation using a male pronoun (28% vs. 18%), while the opposite direction drops the likelihood to 6%. For all traits except Openness, steering significantly increases GLD in one direction while significantly decreasing it in the other ($p < 0.01$). These directions are not consistent across models, indicating different relationships between personality and gender in both models' internal representations. See the full results in the appendix.

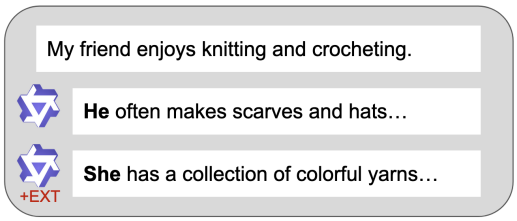

Figure 1: Qwen changes which pronoun it chooses when steered towards an extroverted personality

Conclusions Our results indicate that LLMs' internal representations of complex concepts such as personality and political ideology are entangled with their conception of gender, likely representing learned biases from its training data in ways specific to individual models. Unbalanced datasets of contrastive steering pairs can induce further biases. The biases we observe indicate a need for careful selection of steering datasets and caution when deploying models with steering interventions. The full version of this work will include both larger models and further datasets, including PoliTune (Agiza et al., 2024) and the one used in Gurgurov et al. (2025).

References

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in LLMs. *arXiv preprint arXiv:2402.11190*.

Daniil Gurgurov, Katharina Trinley, Ivan Vykopal, Josef Van Genabith, Simon Ostermann, and Roberto Zamparelli. 2025. Multilingual political views of large language models: Identification and steering. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 279–298.

Petri J Kajonius and John A Johnson. 2019. Assessing the structure of the five factor model of personality (ipip-neo-120) in the public domain. *Europe's Journal of Psychology*, 15(2):260.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*, volume 2025, pages 7180–7211.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Steering towards fairness: Mitigating political stance bias in LLMs. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 52–61, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kenneth JK Ong, Lye Jia Jun, Hieu Minh Nguyen, Seong Hah Cho, Natalia Pérez-Campanero Antolín, and 1 others. 2025. Identifying cooperative personalities in multi-agent contexts through personality steering with representation engineering. *arXiv preprint arXiv:2503.12722*.

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, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

A Appendix: Gender bias results

Model	Steering	GLD	fem%	masc%
Llama	baseline	.080	.394	.168
$\alpha = 1$	left	.720	.794	.000
	right	.350	.010	.934
Qwen	baseline	.048	.044	.629
$\alpha = 2$	left	.069	.657	.005
	right	.139	.003	.745

Table 1: Results for political steering (gender-balanced dataset). All GLD differences to baseline are strongly significant (t-test over all items, $p \ll 0.01$). fem% and masc% are percentage of generated responses containing feminine or masculine pronouns respectively. α is steering strength; Qwen consistently requires higher strengths for equivalent steering performance.

Model	Trait	GLD	fem%	masc%
Llama	baseline	.080	.371	.175
($\alpha = 1$)	EXT+	.047*	.325	.281
	EXT-	.103*	.284	.064
	AGR+	.154*	.298	.082
	AGR-	.045*	.338	**.302**
	CSN+	.059*	.387	.235
	CSN-	.114*	.405	.090
	EST+	**.195***	.454	.057
	EST-	.047*	.193	.340
	OPN+	.058*	.240	.209
	OPN-	.077	**.479**	.160
Qwen	baseline	.048	.044	.629
($\alpha = 2$)	EXT+	.034*	.075	.526
	EXT-	**.063***	.049	**.778**
	AGR+	.062*	.034	.647
	AGR-	.034*	.116	.624
	CSN+	.046	.067	.670
	CSN-	.045	.036	.381
	EST+	.053	.015	.678
	EST-	.042*	**.119**	.580
	OPN+	.044	.034	.312
	OPN-	.050	.041	.562

Table 2: Results for personality steering. Traits (in order) are extroversion, agreeableness, conscientiousness, emotional stability (equivalent to neuroticism scored in reverse), openness. GLD is averaged over all items. An asterisk * indicates highly significant GLD differences compared to the baseline (t-test over all items, $p < 0.01$). Strongest gender preferences and highest GLDs for each model highlighted in **bold**.