Training-free personalisation via *contrastive activation addition* on Llama-3.1-8B.

Steering a frozen language model toward PRISM-preferred behaviour by adding one direction to its residual stream, and the trade-off against generation quality.

AUTHOR
Rares Diaconescu
rdiaconescu@tudelft.nl

EEMCS, Delft University of Technology · The Netherlands

1. Background

- **Value alignment collapse:** Centralised alignment (RLHF and DPO) collapses value pluralism by training a single consensus reward model, erasing minority preferences.
- **Representation steering:** Controls high-level concepts using linear directions in the residual stream, keeping base model weights frozen.
- **Key advantages:** RepE steering introduces zero prompt token overhead, requires no heavy adapters, and allows dynamic, per-forward-pass control.
- **Contrastive activation addition:** CAA extracts steering directions by taking the difference-of-means of activations on preference pairs.

2. Research question

To what extent can representation steering align a frozen LLM with user-preferred toxicity profiles in PRISM without degrading fluency, calibration, or downstream utility?

- → Does the steering effect concentrate on categories where the baseline model disagrees with human preferences?
- → What is the productive steering-strength operating range before model collapse?
- → Can per-user compositional steering align a frozen LLM to individual user preferences?

3. Approach

Identification. Compute a layer-wise steering direction by subtracting average dispreferred activations from preferred activations across 200 PRISM records. Composed per-user directions use dislike-weighted averages or fitted Bradley-Terry coordinates.

$$v_l = \mathbb{E}_{(p,\, y^+)}\,[h_l(p + y^+)] - \mathbb{E}_{(p,\, y^-)}\,[h_l(p + y^-)]$$

200 PRISM pairs · population mean-diff, dislike-weighted aggregate, and SVD-orthogonalised BT ideal-point basis

Control. Add the steering vector directly to the residual stream on layers 16 to 22 during generation. Composed vectors are smaller than population directions due to cancellation, requiring a larger scaling coefficient ($\alpha \in [7.5, 17.5]$ vs. population $\alpha \in [0.3, 0.6]$).

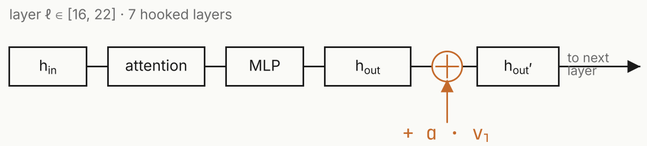

Fig. 1. Intervention on a single model layer. Adding the steering vector alters subsequent layers and tokens, allowing the steering effect to guide the entire generation.

4. Results

Experimental setup. Swept on Llama-3.1-8B (N = 200) under greedy decoding to eliminate temperature-sampling noise. Downstream utility is evaluated via zero-shot MMLU.

Bimodal dose-response. Steering improves toxicity alignment over baseline, showing a non-monotonic relationship. Sweeping the personalized basis reveals a bimodal response curve with two peaks: α = 7.5 (largest relative drop of 14.3% in MAE) and α = 17.5 (lowest absolute toxicity MAE).

RUN / ARM	GPU	A	MAE ↓		PPL ↓	DESCRIPTION
baseline	A100	–	0.01877		4.38	unsteered baseline model
baseline	A10g	–	0.01848		4.47	unsteered baseline model
caa (personalized)	A100	**7.5**	**0.01609**	−14.3%	4.27	**Primary alignment peak (largest relative improvement)**
caa (personalized)	A10g	12.5	0.01660	−10.2%	4.11	Intermediate peak
caa (personalized)	A10g	17.5	0.01601	−13.4%	4.11	Secondary alignment peak (lowest absolute toxicity)
caa (personalized)	A10g	20.0	0.01696	−8.2%	3.92	Optimal text fluency
caa (personalized)	A10g	25.0	0.01853	+0.3%	3.98	Loss of steering effect

Table 1. Sweeping the personalized preference basis under greedy decoding. Relative changes are computed against the respective hardware baseline to isolate floating-point precision shifts.

Selectivity by prompt category. The intervention is highly selective. MAE drops concentrate on categories where the unsteered model disagrees with preferences (harmful_borderline and safe_sensitive), while benign categories are virtually untouched.

category	base A100	base A10g	α = 5.0	α = 7.5	α = 17.5
benign_control	0.00499	0.00470	-7.5%	+4.2%	+5.4%
context_dependent	0.01553	0.01384	+0.6%	-13.8%	-3.8%
harmful_borderline	0.03172	0.03224	-10.6%	-16.3%	-15.7%
safe_sensitive	0.02281	0.02313	-0.4%	-15.8%	-19.6%

% change in MAE vs. baseline · negative = better ≤ 5% 5–15% ≥ 15%

Fig. 2. Selectivity of the steering vector. MAE drops concentrate on the hard categories (up to **−19.6%** drop) while leaving benign prompts essentially untouched. Note: α = 17.5 relative changes are compared to the A10g baseline.

5. Overcoming the quality cliff

Preserving text fluency. Deterministic decoding maintains text quality across the steering range, peaking at alpha = 20.0. This resolves the quality collapse and repetitive loops seen under random sampling, where steering errors accumulate.

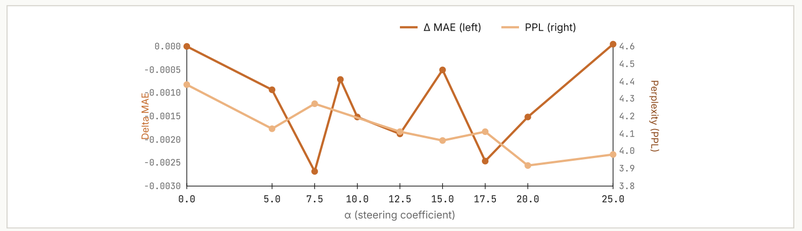

Fig. 3. Delta MAE (orange) and Perplexity (tan) vs. α under greedy decoding. The steering effect shows two peaks at α = 7.5 and 17.5, with perplexity staying flat or improving.

PREFILL ACTIVATION (MDLM)

Gating the steering hook to the first K tokens is a no-op at α = 7.5 (constant MAE ≈ 0.0161), confirming that the steering effect is established during the prefill phase.

6. Discussion & personalisation analysis

Personalization analysis. Safety and user preferences are independent in the dataset, with 96.6% of choices having identical safety ratings. However, on individual user choices, personalized steering achieves the highest accuracy (59.9% versus 54.3% for population steering), demonstrating that personalization successfully matches individual preferences.

Toxicity ties	96.6% of choices have identical safety ratings; only 1.4% differ significantly.	96.6% ties
Chance	Predicting user choices based on safety coordinates alone is equivalent to random guessing (48.3% accuracy).	chance acc
User Match	Personalized steering achieves the best match (59.9% accuracy) when predicting individual user preferences.	59.9% PPA

7. Conclusion

α = 7.5	Primary peak (14.3% error reduction); provides the largest alignment shift.	Primary Peak
α = 17.5	Secondary peak (13.4% error reduction); achieves the lowest toxicity overall.	Secondary Peak
α = 20.0	Highest text fluency (perplexity 3.92); improves response clarity.	best Fluency
α ≥ 25.0	Inactive range; steering effect fades away without model collapse.	inactive
MMLU	Zero-shot reasoning accuracy is fully preserved (67.4% under steering versus 66.8% unsteered).	preserved

8. Future work

- **Sparse features**: Replace dense steering vectors with sparse, interpretable features to isolate individual concepts and prevent unintended changes to other model behaviors.
- **Time-based steering**: Apply steering only during initial prompt reading and the first few generated words to maintain safety while keeping long responses natural.
- **Multi-turn conversations**: Extend testing to multi-step dialogues and reasoning models to study how steering interacts with step-by-step thinking.