STEER-Away: Personalized Safety Alignment via Decoding-Time Logit Steering

Training-free user-specific anti-expert and expert–anti-expert logit steering

Andrei Bogdan Trache atrache@student.tudelft.nl | CSE3000 Research Project | EEMCS, Delft University of Technology

Responsible Professor: Jie Yang | Supervisors: Anne Arzberger, Enrico Liscio

Key result: Training-free logit steering can reduce a personalized toxicity-distance proxy. **Expert–Anti-Expert** gives **+12.65%** mean over four seeds at $\alpha = 2.2$ while keeping MMLU at **70.22%** (base 70.60%) and perplexity near base.

1. Background and motivation

- Safety alignment targets an *average* user, but people disagree on what counts as too toxic – a **pluralistic alignment** problem.
- Per-user fine-tuning is expensive and static.
- **Idea:** keep the model frozen and steer only the next-token logits at decoding time, so the target can change per user.

Contributions:
- User toxicity profiles built from PRISM ratings and Perspective scores.
- Two training-free, decoding-time steering rules on a frozen model.
- A single α-sweep evaluated jointly for safety, utility, and fluency.

2. Research question

Can training-free logit-difference decoding reduce a **user-specific** toxicity-distance, and what does it cost in utility (MMLU) and fluency (perplexity)?

Boundary: "toxicity reduction" = a lower six-dimensional Perspective-score *distance* to the accepted PRISM answer. A proxy metric, not proof that a response is safe.

3. Personalization signal

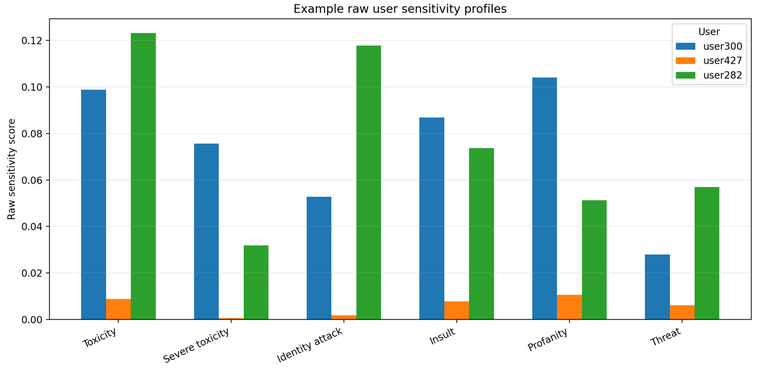
Each user gets one primary category from **PRISM** ratings and **Perspective** scores, chosen from six toxicity dimensions:

`toxicity` `severe` `identity` `insult` `profanity` `threat`

A rated response i gives a dislike weight $d_i = 1 - r_i/100$, and the sensitivity to category c is

$$s_{u,c} = \frac{\sum_i d_i \, P_c(i)}{\sum_i d_i},$$

percentile-normalized across users to pick the steering category $c^*(u)$.



Different users peak on different toxicity dimensions, which is why a single global target loses information.

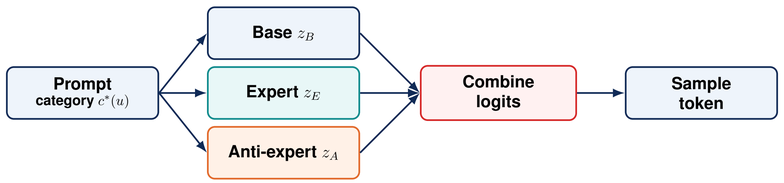
The chosen category $c^*(u)$ is the only user-specific input to steering; the base, expert, and anti-expert branches are otherwise shared.

4. Method: training-free logit steering

A frozen **Llama 3.1 8B** is run in parallel context branches (base, expert, anti-expert). At each step the sampled token is appended to all branches, and their next-token logits are combined before sampling.

Anti-Expert: $z_{\text{new}} = z_B + \alpha(z_B - z_A)$
push away from a category-specific undesired style.

Expert–Anti-Expert: $z_{\text{new}} = z_B + \alpha(z_E - z_A)$
move toward a desired style and away from the undesired one (DExperts-style).

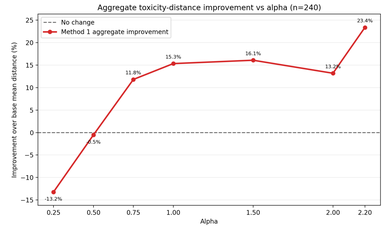


Weights, tokenizer, and base capability stay fixed; α trades intervention strength against distribution shift.

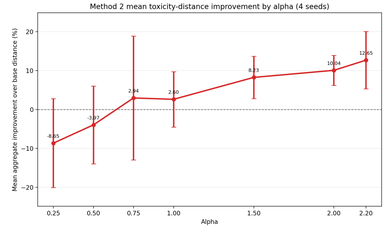
5. Experimental setup

- **Safety:** 240 PRISM prompts per seed (40 per category); distance $D(y, y^*) = \frac{1}{6}\sum_c |P_c(y) - P_c(y^*)|$, reported as aggregate $\%\Delta = 100\,(\overline{D}_{base} - \overline{D}_{method})/\overline{D}_{base}$.
- **Utility:** 1,000 one-shot MMLU questions, macro-averaged over the six steering contexts.
- **Fluency:** generated-answer perplexity scored by the base model.
- **Sweep:** $\alpha \in \{0.25, \ldots, 2.2\}$; temperature 0.7, top-$p$ 0.9, repetition penalty 1.15, 128 tokens. Anti-Expert single-seed; Expert–Anti-Expert over four seeds.

6. Safety results



Anti-Expert, single seed.



Expert–Anti-Expert, mean of four seeds.

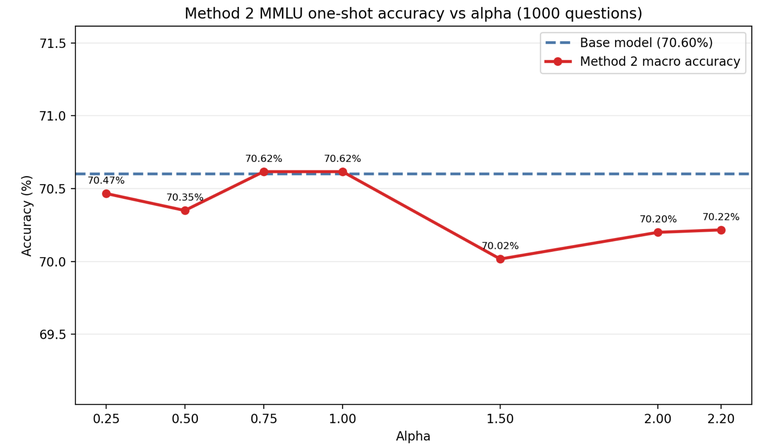
Anti-Expert peak
+23.37%
single seed, $\alpha = 2.2$

Expert–Anti-Expert mean
+12.65%
four seeds (2.11–22.97%)

Both methods help only at **stronger** α: weak steering stays near the base distribution. Anti-Expert reaches the higher single-seed peak, but Expert–Anti-Expert is positive across all four seeds, so it is the more reliable rule.

Gains concentrate at $\alpha \geq 0.75$; below that, steering barely moves the toxicity-distance proxy.

7. Utility and fluency trade-off



Expert–Anti-Expert MMLU stays close to base; Anti-Expert declines as α grows.

Base	Expert–Anti-Expert	Anti-Expert
70.60%	70.22%	68.48%

Expert–Anti-Expert stays close to base MMLU across the whole sweep, while Anti-Expert drops as α grows. Perplexity rises gently (Expert–Anti-Expert: $2.72 \rightarrow 2.96$; Anti-Expert up to 3.32, above base 3.23). Adding the expert direction **preserves utility better** than subtracting an anti-expert alone. Win rate is near 50%, so the gain is aggregate rather than a per-prompt guarantee.

8. Limitations

- Perspective scores and PRISM accepted answers are imperfect proxies / reference points, not ground truth, so a lower distance does not prove a response is safe.
- Each user is reduced to one coarse category; the expert and anti-expert branches are context-induced through prompting, not separately trained models.
- Anti-Expert is reported as a single-seed curve, so its peak may be optimistic; the reported effect is an aggregate mean rather than a per-prompt guarantee.
- Exploratory demographic subgroups (age, gender, religion) are small and imbalanced, and running three branches multiplies the decoding cost at inference time.

9. Conclusions and outlook

- Expert–Anti-Expert logit steering is a promising **lightweight baseline** for personalized toxicity-distance reduction.
- Anti-Expert has the higher peak but a clearer cost; **Expert–Anti-Expert gives the better trade-off and reliability**.
- Unlike trained reward-guided methods (PAD, GenARM), it needs no extra model.
- **Next:** human preference evaluation, per-user or per-category α, larger balanced sub-groups, and trained expert branches.

Key references: DExperts (Liu et al., ACL 2021); Proxy-Tuning (Liu et al., COLM 2024); Decoding-Time Realignment (Liu et al., ICML 2024); PRISM Alignment Dataset (Kirk et al., NeurIPS 2024); Perspective API (Jigsaw); MMLU (Hendrycks et al., ICLR 2021); PAD / GenARM (reward-guided decoding-time personalization).