TU Delft

EEMCS

Value-Aware Post-Decoding Reranking for Training-Free Personalisation of LLM Outputs to User-Specific Toxicity Standards

Iulia Slanina
Responsible Professor: Jie Yang
Supervisors: Anne Arzberger, Enrico Liscio
CSE3000 Research Project

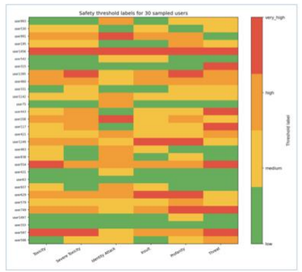
1 Background

LLMs are aligned once, at training time, to a single global notion of toxicity: a universal filter that is costly to train, opaque, and hard to update.

But people disagree on what is toxic. The same sentence is healthy directness to one person and a personal attack to another, and toxicity depends on the utterance, the audience, and the setting.

The PRISM dataset documents this disagreement in real human feedback, and Kirk et al. argue no single standard can satisfy everyone.

Models can be steered at inference time, without retraining: before decoding (prompt), during (logits), or after (reranking). We study the post-decoding stage.



Each row is one PRISM user's toxicity-sensitivity profile across six dimensions. Green = tolerant, red = sensitive. No two rows match.

KEY TAKEAWAY
People disagree on what is toxic. There is no average person to centralise around.

2 Research question

RESEARCH QUESTION
How effectively can value-aware post-decoding align model outputs with user-specific toxicity standards without modifying the underlying generation process?

Gap: prior reranking uses one global scorer, so it cannot tell genuine personalisation from a blanket push toward safer text.

- SQ1 Does it beat an un-steered baseline?
- SQ2 Does an expensive LLM judge beat a cheap matcher?
- SQ3 Real personalisation, or just pick the safest?
- SQ4 What does it cost (fluency, refusals, compute)?

3 Experimental setup

Data.
200 PRISM prompts with a recorded preferred response, balanced over four categories (harmful-borderline, safe-sensitive, context-dependent, benign-control), each matched to its own author's profile. Run on 4 matched seeds = 800 records per scorer.

User profile = dislike-weighted mean toxicity per dimension d.

$$d_p = 1 - R_p / 100$$

$$w_d = \frac{\sum_p d_p \cdot \mathrm{tox}_d(p)}{\sum_p d_p}$$

d = dislike weight, R = rating, tox = Detoxify toxicity score.

A high profile value on a dimension means the user repeatedly disliked toxic content there, so we read higher values as more sensitive (lower = more tolerant).

Evaluation.

- For each prompt, draw N = 8 candidates from LLaMA-3.1-8B; each scorer picks one. Score = MAE to the author's preferred PRISM response over six toxicity dimensions.
- MAE-to-preferred is two-sided: it penalises being too toxic and not toxic enough, so always picking the safest candidate does not win.
- Select with Detoxify, evaluate with Perspective (independent detectors). Same pipeline as the baseline, so all scorers are comparable.

4 Method: four scorers, two families

Each scorer picks one of the 8 candidates for the user. The per-user target above is shared; the families differ in what they see.

- S1 GPT + S4 Claude judges: read the profile + 8 candidates, return a listwise ranking as JSON. Paid, about 30 s per prompt.
- S2 + S3 geometric matchers: score candidates with Detoxify, pick the closest to the target. No LLM, $0, under 1 s.

Shared per-user target and weights

$$\mathrm{target}_d = 100 - \mathrm{pct}_d \qquad w_d = \mathrm{pct}_d$$

pct is the user's sensitivity percentile per dimension. Sensitive dimensions get a low-toxicity target and a large weight.

S2 sensitivity-weighted L1

$$\mathrm{pick} = \arg\min \sum_d w_d \,|\,\mathrm{cand}_d - \mathrm{target}_d\,|$$

Treats the six toxicity axes as independent.

S3 Mahalanobis (Ledoit-Wolf shrinkage)

$$\mathrm{dist} = \delta^\top \Sigma^{-1} \delta \qquad \delta = \mathrm{cand} - \mathrm{target}$$

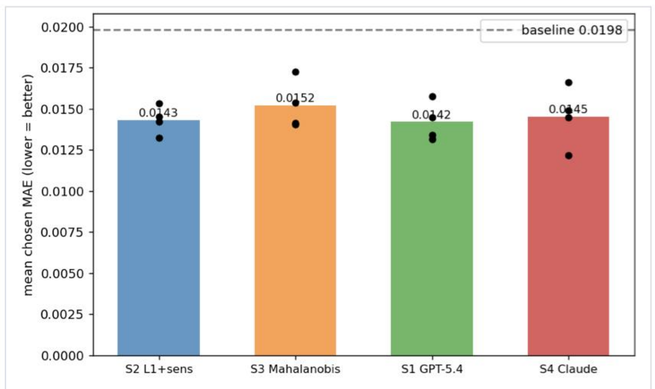
The inverse covariance decorrelates the axes, so correlated movement (toxicity, insult, profanity) is counted once. Ledoit-Wolf gives a stable inverse from 8 candidates.

5 Results

Q: beat the baseline (SQ1), and does the expensive judge win (SQ2)?

A: all four cut per-record error by 23 to 28% and tie at the top. No module wins (every pairwise gap is below the Bonferroni threshold).

Scorer	MAE	vs base	cost
GPT judge	0.0142	−28%	$, 30 s
weighted L1	0.0143	−28%	free
Claude judge	0.0145	−27%	$, 30 s
Mahalanobis	0.0152	−23%	free
baseline	0.0198	n/a	n/a



Mean chosen MAE per scorer (bars), per-seed runs (dots), baseline (dashed). All four sit well below baseline and within 0.001 of each other.

Gains are largest on harmful-borderline prompts (−29 to −37%), where the candidate pool varies most, and fade on benign prompts with no head-room.
Calibration: every scorer picks slightly safer than the preferred level, most on toxicity and identity-attack, so an aggregate metric quietly rewards under-shooting.

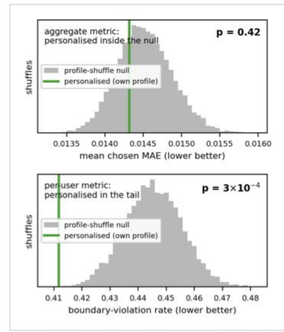
KEY TAKEAWAY
A free, no-LLM matcher ties two frontier judges. Personalisation needs a per-user target, not an expensive scorer.

6 Personalisation

Q: does selection react to the individual, or just pick the safest (SQ3)?

A: yes, it tracks the person. Three checks:

- Cuts land on each person's sensitive dimensions (correlation +0.51 vs +0.2 for a stranger's profile, p < 0.001).
- Remove the per-user weighting and fit gets significantly worse (L1 0.0143 to 0.0158, Mahalanobis 0.0152 to 0.0190, p < 0.001).
- Swap in someone else's profile and 41% of judge picks change.



Same profile-shuffle test, two metrics. Aggregate mean MAE cannot see the per-user effect (top, p = 0.42); a per-user boundary-violation rate can (bottom, p = 0.0003).

KEY TAKEAWAY
The per-user signal is real, but only a per-user-sensitive metric reveals it. Situated alignment needs a better ruler, not more steering.

7 Costs and trade-offs

Q: what does personalisation cost (SQ4)? A: very little, except the judges' paid calls.

	judges	matchers
perplexity	7.2-7.6	7.6
refusal	~1%	~1%
time / prompt	~30 s	< 1 s
API cost / run	$15-25	$0

Baseline perplexity 6.38, refusal ~1%: fluency drops under one unit and refusals are unchanged.
The MMLU utility loss comes from sampling diverse candidates, not from the value-aware selection.

KEY TAKEAWAY
The cheap geometric matcher is the sensible default; reserve the LLM judge for when its text-level reasoning earns the spend.

8 Future work & limitations

Limitations

- Profiles need 20+ ratings per user (active-user bias).
- Profile is inferred from a biased detector, not asked of users.
- Judge APIs force temperature 1, so judges are not fully reproducible.
- One generator (LLaMA-3.1-8B) and one dataset (PRISM).

Future work

- Per-user-sensitive metrics as the default (the boundary-violation rate).
- Larger pools (N > 8) and tuned sampling temperature.
- Repeat with a different detector; elicit profiles directly from users.