Personalizing Toxicity Sensitivity without Retraining

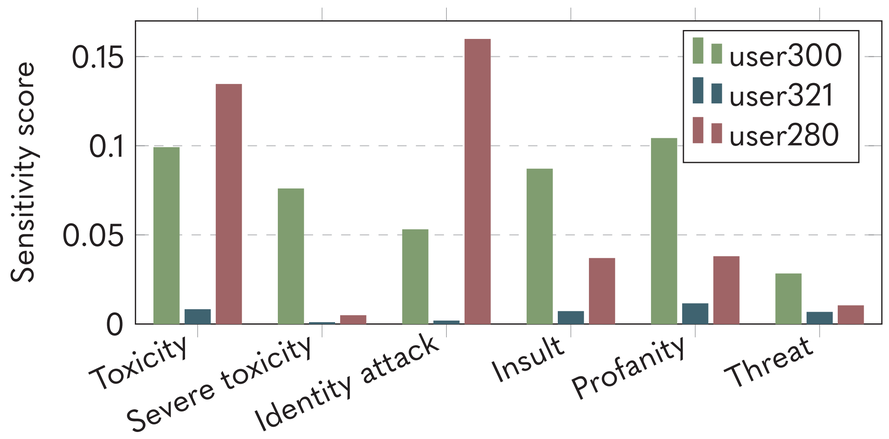
Sensitivity is Personal

What counts as "toxic" isn't universal.

A slur for you is a reclaimed term to another. Graphic medical details can upset one and reassure another.

When LLMs are aligned to a single notion of toxicity, **the threshold may be too strict for some users and not strict enough for others**.

Toxicity Sensitivity Differs across the PRISM Dataset



For each PRISM user, we score every response they rated with Perspective API.

Each response's scores are weighted by how much the user disliked it, so user's sensitivity is highest in the categories they consistently rated down.

Training-free Personalization Methods

Pre-Decoding (←)

URIAL: add user-specific safety instructions and examples before the prompt.

PBPO: a separate model rewrites the prompt to neutralize wording the user is sensitive to.

In-Decoding (↓)

PCAA (activation steering): steer the model's hidden states along a direction personalized to the user.

CGD: a toxicity classifier, weighted by the user's profile, penalizes candidate tokens at every step.

EADS: steer each token toward an *expert* branch and away from an *anti-expert* branch.

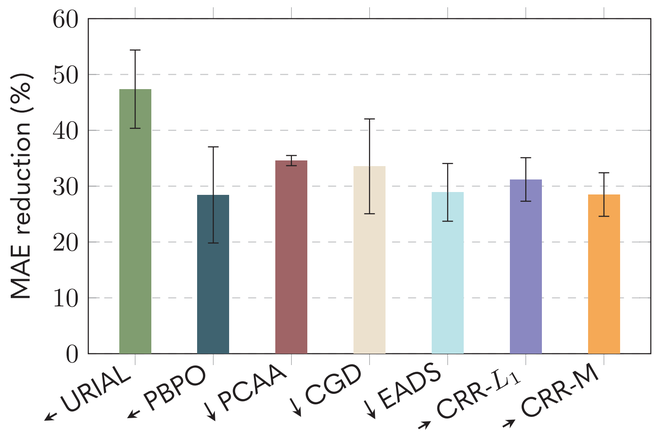
Post-Decoding (→)

CRR-L_1: generate several candidate responses and pick the one closest to the user profile.

CRR-M: same as CRR-L_1, but the Mahalanobis distance accounts for correlation between categories.

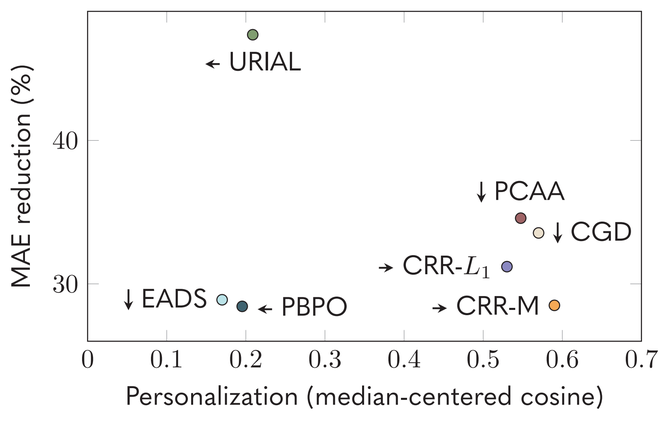
Overall Alignment ...

All methods gets closer to the users' sensitivity profiles, cutting error by 28-47%.



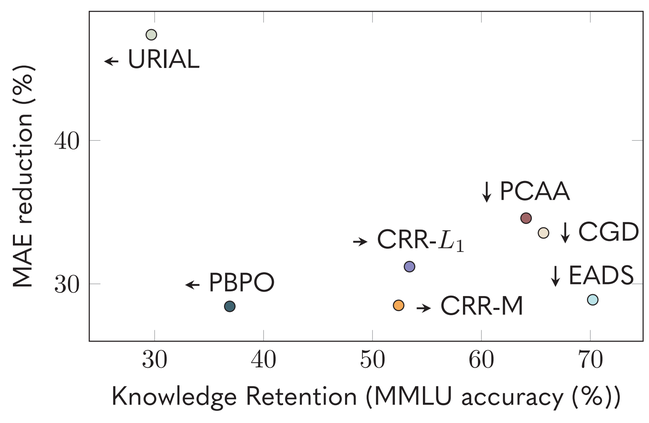
... vs. Personalization

Lower error ≠ personalization (URIAL just makes everything safer).



... vs. Knowledge Retention

Sensitivity alignment can trade off with knowledge retention (and fluency).



Takeaways

No method wins on everything.
All methods improve alignment, but the ranking changes once you factor in knowledge retention and fluency.

Lower error ≠ personalization.
Some methods improve alignment by making text safer for everyone rather than adapting to each user's sensitivities.

Gains are uneven across groups.
The best method differs by demographic groups, and one method even worsens alignment for one group.

R.A.C. Diaconescu*, I. Slănină*, A. Florea*, A.B. Trache*, M.E. Coroi*, A. Arzberger, J. Yang, E. Liscio.

"Beyond a Global Norm: Personalizing Toxicity Sensitivity in Language Models Without Retraining". WOAH @ EMNLP2026.

TU Delft
Hybrid Intelligence
AlgoSoc /