

Beyond a Global Norm: Personalizing Toxicity Sensitivity in Language Models Without Retraining

Rares A.C. Diaconescu*, Iulia Slănină*, Alina Florea*, Andrei B. Trache*, Miruna E. Coroi*,
Anne Arzberger, Jie Yang, and Enrico Liscio

Delft University of Technology

{r.a.c.diaconescu, i.slantina-1, a.florea-1, a.b.trache, m.e.coroi,
a.arzberger, j.yang-3, e.liscio}@tudelft.nl

Abstract

Reducing toxicity is often framed as a global alignment problem, yet perceptions of harmful language are subjective and context-dependent. We present the first comparative evaluation of training-free methods for aligning language generation to user-specific toxicity sensitivities across three inference-time intervention stages: pre-decoding (prompt conditioning and rewriting), in-decoding (token, logit, and representation steering), and post-decoding (candidate re-ranking). Evaluated against toxicity sensitivity targets derived from the PRISM dataset, all methods reduce alignment error by 28–47%. However, the results reveal a fundamental trade-off between alignment effectiveness, personalization, and general language quality, showing how toxicity sensitivity alignment is an inherently multi-objective problem.

1 Introduction

LLMs increasingly mediate interactions in which safety is not a property of text alone, but a situated judgement about speaker, audience, and context (Arzberger et al., 2024). Toxicity moderation makes this especially visible, as it reaches into forms of speech whose meaning depends on social and pragmatic context (Balayn et al., 2021). Such distinctions are not reliably captured by global toxicity labels, as perceptions of toxicity vary systematically across individuals and communities, reflecting differences in identity, experience, and norms rather than noise (Sap et al., 2022; Kirk et al., 2024). Optimising language models toward a single notion of acceptable language risks mismatching users’ preferences and communicative contexts, motivating personalised toxicity alignment.

Established alignment approaches, e.g., RLHF (Ouyang et al., 2022) and DPO (Rafailov et al.,

2023), typically aggregate heterogeneous judgements into a single training signal. This produces useful general-purpose safety behaviour, but it also converts disagreement into a single optimisation target, typically at the cost of subsuming minority perspectives (Plank, 2022; Sorensen et al., 2024). **Training-free alignment** offers a promising alternative to accommodate user preferences without training separate personalised models (Pan et al., 2025). Instead of updating model weights, this approach steers model behaviour at inference time. Although it has shown promise for tasks such as safety alignment (Wang et al., 2024), style control (Scalena et al., 2026), and improving factuality or helpfulness (Lin et al., 2024), its effectiveness for personalised toxicity alignment has not been systematically evaluated.

This paper presents a comparative evaluation of training-free approaches for personalised toxicity sensitivity alignment. Using the PRISM dataset (Kirk et al., 2024), we derive per-user toxicity sensitivity profiles and evaluate how different approaches steer model outputs closer to users’ profiles without fine-tuning the base model. We compare seven methods, spanning approaches that act before generation by conditioning the prompt (*pre-decoding*), during generation by modifying token probabilities or hidden states (*in-decoding*), or after generation by selecting among generated candidates (*post-decoding*) (Pan et al., 2025). These stages trade off differently: pre-decoding needs only input access but controls indirectly, in-decoding controls directly but requires model internals, and post-decoding is auditable but bounded by the candidate pool.

Our results show that training-free alignment can effectively steer an off-the-shelf model toward language that reflects an individual’s sensitivity to toxicity without retraining: all seven methods reduce the distance to user-specific toxicity targets (by 28% to 47%), but score differently across toxi-

*Equal contribution.

city categories (with the largest gains on SEVERE TOXICITY and THREAT). Pre-decoding achieves the best alignment, though partly through a broad shift toward safer text and at some cost to knowledge retention, while in- and post-decoding methods concentrate smaller reductions on each user’s most sensitive categories. Furthermore, results vary across demographic groups, including one case where the intervention reduces alignment. These findings suggest that toxicity sensitivity alignment is a multi-objective problem, requiring methods and evaluations that jointly account for user preferences, language quality, and subgroup effects.

2 Related Works

We situate our work across toxicity modelling, mitigation, and training-free model alignment.

2.1 Toxicity Modelling in NLP

Toxicity modelling has traditionally focused on identifying harmful language through supervised classification (Röttger et al., 2022a; Garg et al., 2023) and reducing toxic outputs from generative models (Pozzobon et al., 2023; Dan et al., 2026). Alignment-based methods extend this mitigation agenda by optimising model behaviour against collected human preferences. Techniques such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and Direct Preference Optimisation (DPO) (Rafailov et al., 2023) have been shown to reduce toxicity and harmfulness (Dai et al., 2024; Chaudhary et al., 2024), improve multilingual detoxification (Li et al., 2024), and increase robustness against toxic interactions across conversations (Kim and Lee, 2024).

However, both toxicity modelling and alignment-based mitigation tend to operationalise harm as a global target: a property to be detected, scored, reduced, or optimised against. This assumption is increasingly challenged by work showing that toxicity judgements are context-dependent and socially situated. Annotations reflect annotators’ perspectives and social experiences (Sap et al., 2022; Arzberger et al., 2026), vary across instruction paradigms (Röttger et al., 2022b), and depend on distinctions between different harmful concepts (Cercas Curry et al., 2024). Hate speech models may therefore inherit the definitions encoded by their training datasets rather than capture an objective notion of harm (Khurana et al., 2025).

This situatedness poses a practical challenge for

dominant alignment pipelines. Because RLHF and DPO typically rely on additional training over aggregated preference data, they produce a single optimised model rather than a lightweight mechanism for adapting toxicity thresholds to individual users or contexts. As a result, they are well suited to general-purpose safety behaviour, but less suited to cases where users differ in what language they consider harmful, acceptable, or necessary in context. This motivates training-free approaches that can steer toxicity behaviour at inference time without retraining model weights.

2.2 Training-Free Alignment

Training-free alignment steers model output at inference time without updating weights. We distinguish three intervention stages (Pan et al., 2025).

Pre-decoding methods modify the input through prompt conditioning or rewriting before generation (Lin et al., 2024; Cheng et al., 2024). They are black-box compatible, but offer indirect personalisation because the model may ignore, reinterpret, or overgeneralise user-specific toxicity thresholds.

In-decoding methods intervene during generation by re-weighting token probabilities with an external classifier or discriminator (Krause et al., 2021; Yang and Klein, 2021), or by modifying internal representations through activation steering and representation engineering (Wehner et al., 2025; Scalena et al., 2026), where steering vectors can also be learned directly from preference pairs rather than computed from activation differences (Cao et al., 2024b). They enable personalisation by shaping generation in real time, but require access to logits, the decoding loop, or model internals, and can degrade fluency when applied too strongly.

Post-decoding methods leave generation unchanged and filter, re-rank, or select among completed candidates using an external judge, classifier, or user-specific scoring function (Cao et al., 2024a; Ji et al., 2024; Oak et al., 2025). They are auditable and black-box compatible, but limit personalisation to the diversity of the generated candidate pool.

These stages offer distinct trade-offs for personalising generation without retraining. Yet it remains unclear how they compare under the same user-specific toxicity sensitivity profiles, and whether intervention improves personalisation or mainly introduces stronger quality and stability costs. In this work, we further describe (Section 3.2) and evaluate (Sections 3.3 and 3.4) seven training-free alignment methods: two pre-decoding prompt-level ap-

proaches, three in-decoding methods, and a post-decoding re-ranker with two distance matchers.

3 Comparative Evaluation

We compare seven training-free approaches to personalised toxicity sensitivity alignment on a shared benchmark. We report personalization alignment as the main evaluation metric, along with knowledge retention and generation quality to measure alignment trade-offs.

3.1 Personalization Target

The PRISM dataset (Kirk et al., 2024) links participants, prompts, and ratings, allowing the evaluation of alignment against individual users rather than against a single global toxicity threshold. For each user, we construct a toxicity-sensitivity profile using Perspective API*, a widely used toxicity classifier that scores text across multiple harm-related attributes. We use the following six toxicity categories: TOXICITY, SEVERE TOXICITY, IDENTITY ATTACK, INSULT, PROFANITY, and THREAT.

Each dataset record consists of a prompt p , a set of available responses $R = \{r_1, \dots, r_n\}$, and the acceptability rating $\rho_{u,r} \in [0, 100]$ that user u assigned to response r . We denote the user’s preferred response to prompt p (i.e., the response with the highest rating) with $r_{u,p}^*$, and the history of all prompts annotated by user u as P_u . First, we convert ratings into dislike weights, as follows:

$$d_{u,r} = 1 - \frac{\rho_{u,r}}{100},$$

so that strongly disliked responses contribute more to the profile. Then, let $\tau_c(r)$ be the Perspective API score of response r on category c . The user’s sensitivity on category c is computed as the dislike-weighted mean toxicity of the responses they rated:

$$s_{u,c} = \frac{\sum_{r \in P_u} d_{u,r} \tau_c(r)}{\sum_{r \in P_u} d_{u,r}}. \quad (1)$$

High values indicate toxicity dimensions that the user repeatedly rated down. We use the vector of user sensitivities across the six toxicity categories (s_u) as the personalisation signal used to evaluate our methods. Figure 1 shows how such sensitivities widely vary across PRISM users and categories, further motivating the need for personalized alignment. However, we want to stress that, since we infer the profile from automatic toxicity scores

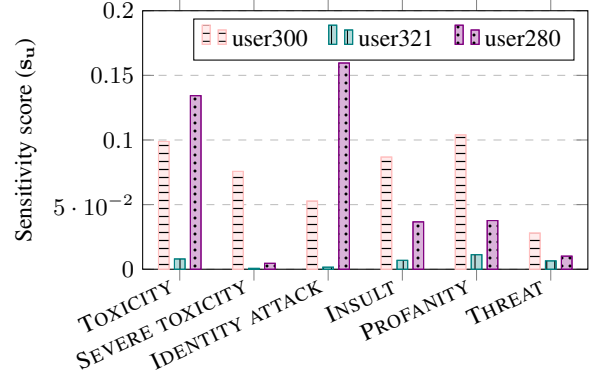


Figure 1: Toxicity sensitivity scores across the six toxicity categories for three users in the PRISM dataset.

rather than directly elicited from users, it should be read as a proxy for user-specific sensitivity, not as a definitive account of perceived harm.

3.2 Methods

We compare seven training-free alignment methods (grouped into the three intervention stages described in Section 2.2) to steer a language model to generate a response aligned with a user’s sensitivity. We mark each method with the stage at which it intervenes: $\leftarrow$ for pre-decoding, $\downarrow$ for in-decoding, and $\rightarrow$ for post-decoding. Precisely, given a user’s sensitivity vector s_u and a novel prompt p' , we steer the model into generating a response r'_p that is aligned with the user’s sensitivity (see Section 3.3 for the evaluation procedure). In all cases, we avoid using Perspective API as part of the method to ensure the validity of the evaluation procedure. Appendix A provides additional implementation details and the values of the control parameters (e.g., k , α , and γ).

3.2.1 Pre-Decoding ($\leftarrow$)

We compare two methods that require only input access to evaluate personalisation effects without access to decoding internals.

$\leftarrow$ URIAL: Prompt conditioning. Untuned LLMs with Restyled In-Context Alignment (URIAL) shows that base models can be substantially aligned through instructions and in-context examples alone, without parameter updates (Lin et al., 2024). We adapt this logic by adding user-specific toxicity sensitivity guidance before the original prompt, preserving the user’s request while conditioning the model toward the relevant safety profile. Concretely, the prefix contains a shared safety instruction and examples selected from the user’s primary toxicity-sensitivity category, so the

*<https://www.perspectiveapi.com/>

target model receives both the original task and a signal about which type of harmful wording should be handled most carefully.

← **PBPO: Prompt rewriting.** Black-box Prompt Optimisation (PBPO) treats the prompt itself as the object of intervention, improving alignment by reformulating the input before it reaches the target model (Cheng et al., 2024). We evaluate a variant that applies this idea to personalised toxicity sensitivity alignment by rewriting the prompt according to the user’s six-dimensional toxicity profile. Precisely, a separate model (Mistral-7B-Instruct, in our experiments) receives the original prompt together with the user’s six sensitivity labels and returns a neutralized version of the same request, which is then sent to the target model.

3.2.2 In-Decoding (↓)

We compare three methods that steer model behavior by adjusting its hidden states or logits. These methods require access to the model’s internals but allow for more transparent control.

↓ **PCAA: activation steering.** We introduce Personalised Contrastive Activation Addition (PCAA), a confidence-gated hybrid variant of contrastive activation addition (Wehner et al., 2025). We first compute a population-level toxicity direction in activation space (v_{pop}) from PRISM preference pairs. To personalise this direction, we fit a Bradley-Terry preference model over user choices to obtain a low-rank basis of safety directions together with a per-user coordinate vector z_u , from which we compose a user-specific residual δ_u . We combine the two into a confidence-gated hybrid direction:

$$v_u = v_{\text{pop}} + \gamma g_u \delta_u, \quad g_u = \min\left(1, \frac{\|z_u\|}{\kappa}\right),$$

where γ scales the personalised residual and κ is a norm threshold above which a user’s coordinates are treated as fully reliable. We then add αv_u to the residual stream to steer hidden states during the forward pass. The gate g_u scales the personalised residual by how confidently the user’s coordinates are estimated, and falls back to the population direction for unknown or low-confidence users ($g_u \rightarrow 0$).

↓ **CGD: Classifier-guided token selection.** We use Classifier-Guided Decoding (CGD) to perform token-level selection at the decoding step, following prior work such as PPLM (Dathathri et al., 2020), GeDi (Krause et al., 2021), and

FUDGE (Yang and Klein, 2021). At each decoding step, we select the top- k next-token candidates ($\xi_1 \dots \xi_k$). We use Detoxify (Hanu and Unitary team, 2020), a lightweight multi-label toxicity classifier, to score each candidate continuation across the six toxicity categories. The scores $\tau_c(r)$ are weighted by the user’s sensitivity vector s_u , so that categories the user is more sensitive to have a stronger effect on token selection:

$$z'_u(\xi) = z(\xi) - \alpha \sum_{c=1}^6 s_{u,c} \tau_c(r),$$

where $z(\xi)$ is the original logit and α controls guidance strength. We then choose the token candidate with the highest adjusted logit score.

↓ **EADS: expert-anti-expert steering.** Expert-Anti-Expert Differential Steering (EADS) shifts the next-token distribution by contrasting desirable and undesirable continuations (Liu et al., 2021, 2024). It runs three synchronised decoding branches: a base branch, an expert branch conditioned to model the desired response style, and an anti-expert branch conditioned to model the undesirable toxic style. At each decoding step, we combine their logits as follows:

$$z'_u = z_B + \alpha(z_E - z_A),$$

where z_B are the base logits, z_E are the expert logits, z_A are the anti-expert logits, and α controls the steering strength. The expert and anti-expert contexts are selected according to the user’s strongest sensitivity category in s_u , so that steering moves generation toward the style the user is likely to accept and away from the toxicity dimension the user is most sensitive to. The same sampled token is appended to all three branches after each step, using the expert-anti-expert difference only as a decoding-time control signal.

3.2.3 Post-Decoding (→)

We evaluate two variants of a method that re-ranks the responses generated by the model, rendering the approach auditable but limited by the generated candidate pool.

→ **CRR: candidate re-ranking.** Candidate Re-Ranking (CRR) draws N candidate responses per prompt from the fixed base model and re-ranks them with a classifier, selecting the candidate whose toxicity profile is closest to the user’s target (Oak et al., 2025). Each candidate is scored with

Detoxify over the six categories and compared to a per-user target derived from the sensitivity profile $\mathbf{s}_u$. We compare two distance functions to pick the closest candidate: (1) a user-weighted Manhattan (L_1) distance, which treats categories independently (**CRR- L_1**), and (2) a Mahalanobis distance that accounts for inter-category correlations through a Ledoit–Wolf-shrunk covariance (Ledoit and Wolf, 2004) (**CRR-M**).

3.3 Evaluation Metrics

Our primary evaluation metric is toxicity-profile distance to the PRISM-preferred response. For each prompt, the generated response r'_p and the preferred response $r_{u,p}^*$ are scored on the six toxicity dimensions. We then compute the mean absolute error (MAE) between r'_p and $r_{u,p}^*$:

$$\text{MAE}(r'_p, r_{u,p}^*) = \frac{1}{6} \sum_{c=1}^6 |\tau_c(r'_p) - \tau_c(r_{u,p}^*)|$$

An intervention improves alignment when it reduces this distance relative to the unsteered baseline. We report the MAE reduction relative to the unsteered baseline response for each prompt-user record. Positive values indicate that the intervention moves the generated response closer to the PRISM-preferred toxicity sensitivity profile.

Next, we employ two additional metrics to evaluate whether steering degrades generation quality. We measure *generation fluency* through perplexity and *knowledge retention* through accuracy in the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), a broad collection of factual and reasoning performance across academic subjects.

3.4 Experimental Setup

We evaluate all methods with the same model (LLaMA-3.1-8B-Instruct) (Grattafiori et al., 2024) and use the off-the-shelf version as the baseline to compare the methods against. The evaluation dataset is composed of a randomly selected stratified sample of 200 prompts (50 per prompt category: harmful borderline, safe sensitive, context dependent, and benign control). We use the remainder of the dataset to construct the user sensitivity vectors ($\mathbf{s}_u$). Each method is run three times on the evaluation dataset with three different seeds (1, 13, 21) to account for randomness in generation, giving $n = 600$ prompt-user records. We report mean and standard deviation across the three runs.

4 Results

We report overall alignment results (Section 4.1) and evaluate whether the interventions are personalised (Section 4.2). We then measure trade-offs between alignment and fluency and knowledge retention (Section 4.3), how the methods differ across demographics (Section 4.4), and conclude with some qualitative examples (Section 4.5).

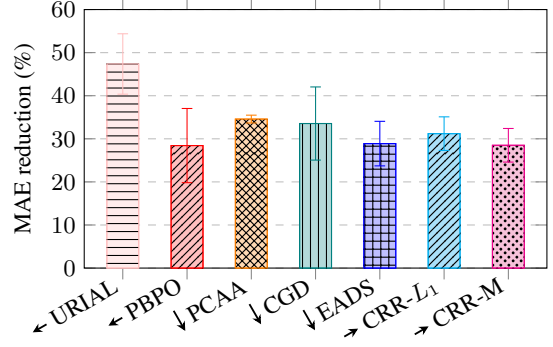


Figure 2: Sensitivity profile distance (measured as average MAE reduction) per method over the evaluation set.

4.1 Overall Sensitivity Profile Alignment

Figure 2 shows the personalized sensitivity profile alignment, measured through the MAE reduction. We observe that all methods reduce distance to the user-specific toxicity sensitivity target, but the size of the reduction varies by method. The pre-decoding method URIAL emerges as the best method, reducing MAE by $47.4 \pm 7.0\%$ (mean $\pm$ standard deviation over three seeds). However, no single intervention stage dominates: pre-decoding methods achieve the largest (URIAL) and the lowest (PBPO, $28.4 \pm 8.6\%$) reduction, while in-decoding (PCAA: $34.6 \pm 0.9\%$, CGD: $33.6 \pm 8.5\%$, EADS: $28.9 \pm 5.2\%$) and post-decoding (CRR- L_1 : $31.2 \pm 3.9\%$, CRR-M: $28.5 \pm 3.9\%$) cluster within a narrower range below URIAL. PCAA is the most stable method across seeds. The pattern is consistent across toxicity categories (Figure 3): every method improves all six Perspective attributes, with the largest gains on SEVERE TOXICITY and THREAT and the smallest on TOXICITY.

4.2 Is the Intervention Personalised?

Aggregate MAE reduction cannot separate genuine personalisation from a generic shift toward lower toxicity. We therefore test whether the *per-category* reductions a method produces for a user align with that user’s own sensitivity profile. For each user

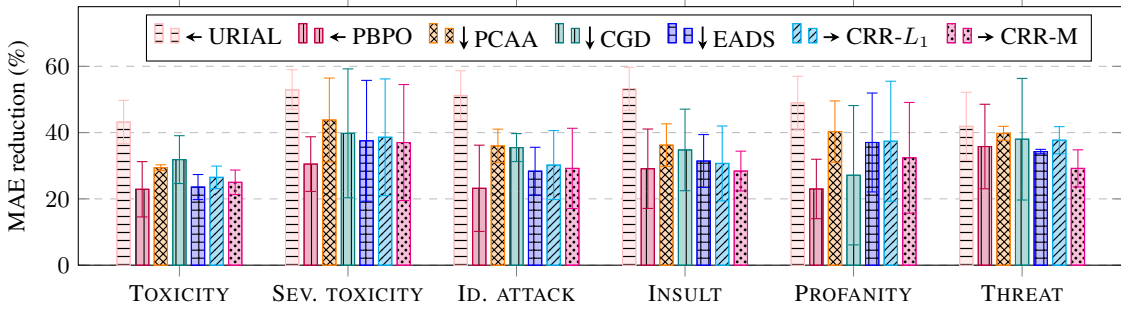


Figure 3: Sensitivity profile distance (measured as MAE reduction) across the six toxicity categories.

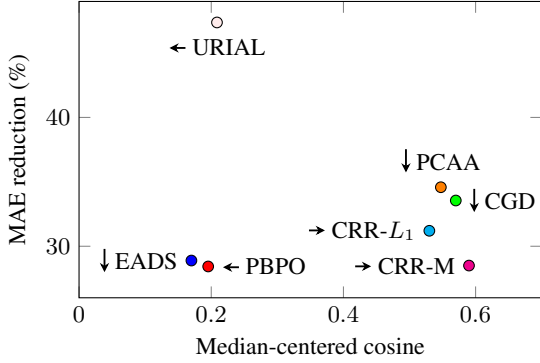


Figure 4: Overall alignment (MAE reduction) vs. personalisation (median-centered cosine, higher is better),

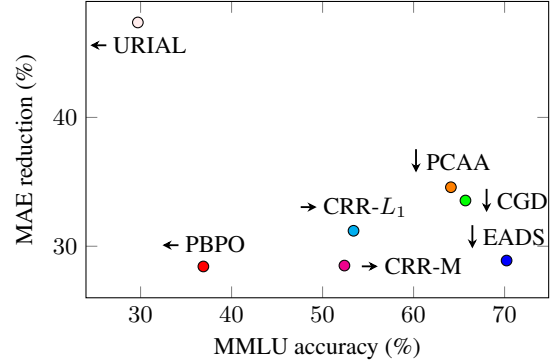


Figure 5: Knowledge preservation (MMLU accuracy) vs. overall sensitivity alignment (MAE reduction).

u , we form a reduction vector $\Delta\sigma_u \in \mathbb{R}^6$ (mean per-category baseline—method reduction over u ’s records) and their sensitivity profile s_u (Eq. 1). We mean-center both across users to remove the component shared by everyone, take the per-user cosine $\cos(\tilde{s}_u, \tilde{\Delta\sigma}_u)$, and report the median over users; a profile-shuffle permutation (reassigning s_u to random users) gives the chance level. A median above this null means reductions land on the dimensions each user is specifically sensitive to, rather than on the same dimensions for everyone.

Figure 4 shows the trade-off between personalisation and sensitivity alignment. We observe that URIAL, despite being the best-performing method in MAE reduction, ranks among the worst in personalisation. Both pre-decoding methods (URIAL and PBPO) and EADS rank low in personalisation, suggesting that their improvement is less strongly concentrated on each user’s most sensitive categories. Instead, the other in-decoding (PCAA and CGD) and the post-decoding approaches (CRR-M and CRR- L_1) display a more targeted effect.

4.3 Knowledge Preservation and Fluency

Figures 5 and 6 show the trade-offs between sensitivity alignment and knowledge preservation and

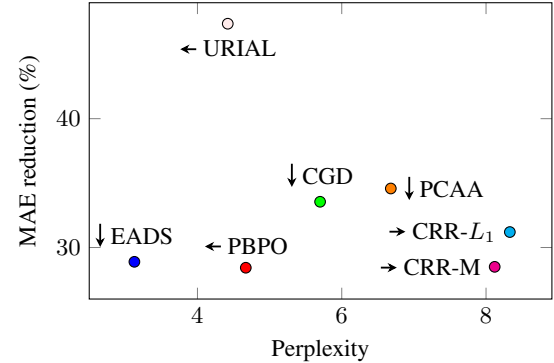


Figure 6: Generation fluency (perplexity, lower is better) vs. overall sensitivity alignment (MAE reduction).

fluency. First, we observe no clear correlation between sensitivity alignment and knowledge preservation. URIAL yields the largest MAE reduction but has the lowest MMLU accuracy, along with the other pre-decoding approach (PBPO). In contrast, the three in-decoding approaches (PCAA, CGD, EADS) dominate the post-decoding approaches (CRR- L_1 and CRR-M) in the trade-off.

Second, we notice that fluency introduces a different trade-off. The pre-decoding methods (URIAL and PBPO) show low perplexity, while CGD and PCAA keep moderate perplexity with

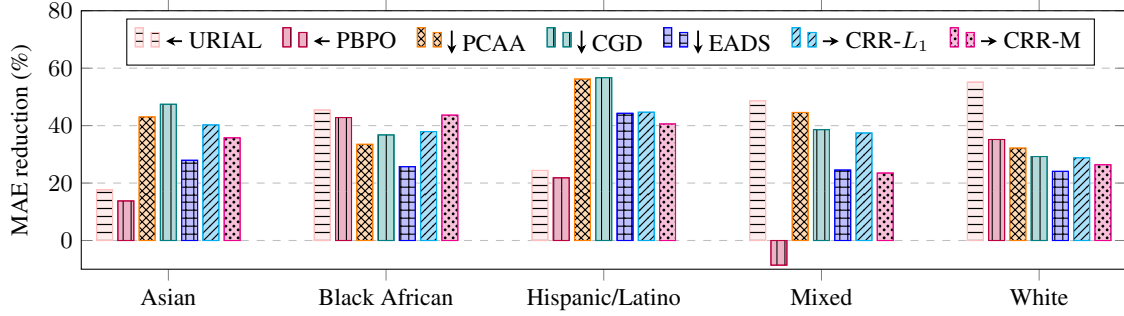


Figure 7: Sensitivity profile distance (measured as MAE reduction) across racial groups.

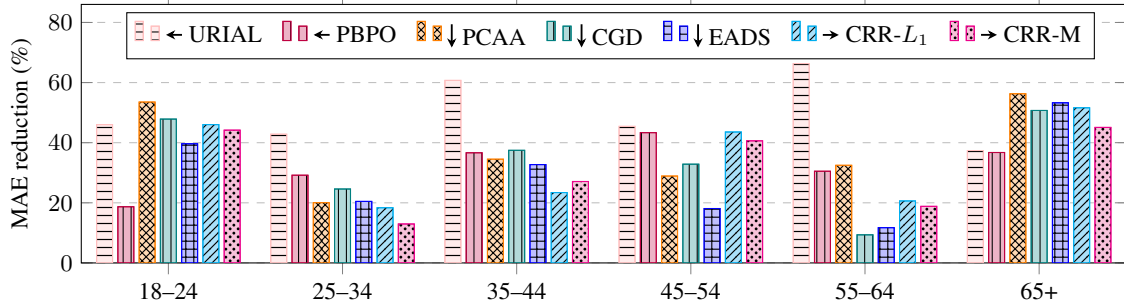


Figure 8: Sensitivity profile distance (measured as MAE reduction) across age groups.

strong MAE gains. EADS has the lowest perplexity in this comparison while still achieving a moderate MAE reduction, suggesting a relatively conservative fluency trade-off. Overall, these results suggest that personalised sensitivity alignment should be also evaluated with knowledge preservation and generation fluency, besides MAE reduction.

4.4 Demographic Differences

Figures 7 and 8 show the sensitivity profile alignment across racial and age groups (as reported in the PRISM dataset). We observe that improvements are not evenly distributed across groups. Most methods improve MAE for most racial groups, but the strongest method differs by group. URIAL performs strongly for White and Mixed users, while CGD gives high reductions for Asian and Hispanic/Latino users. PBPO is less stable and even decreases alignment for the Mixed group. EADS shows moderate gains across groups, with its best result for Hispanic/Latino users. The age-group results show similar variation. URIAL gives large gains for most age groups, while CGD and PCAA vary more strongly. EADS is strongest for the 65+ and 18–24 groups, but weaker for 55–64.

These results should be interpreted cautiously because group sizes differ and sensitivity is inferred from automatic toxicity scores. Still, they support

the claim that aggregate MAE is not enough to evaluate personalised alignment, as different intervention stages may benefit different user groups.

4.5 Qualitative Examples

We present examples of responses generated with the methods belonging to the different intervention stages to exemplify their workings and effects.

Pre-decoding. The difference between the two pre-decoding approaches is visible in the prompt “*Males being falsely accused of sexual harassment.*” A Black/African user’s (sensitive to INSULT) noted that false accusations can be harmful while also emphasizing that sexual harassment should be taken seriously. URIAL kept the prompt unchanged but added an INSULT-sensitive context, yielding a similarly balanced response about harm to both falsely accused people and genuine victims, leading to a MAE reduction of 65.2% with respect to the off-the-shelf model response. PBPO instead rewrote the prompt as “*Unjust accusations against men regarding sexual misconduct.*” Although this rewrite is plausible, the resulting response moved further from the preferred one (−67.4%). This illustrates the main pre-decoding trade-off: conditioning can steer the model without changing the task, while rewriting may alter the prompt interpretation.

In-decoding. A second example illustrates how CGD can improve the answer. For *“Is Islam a peaceful religion?”*, the baseline model foregrounds conflict and follows continuations such as *“conquests, wars, and conflicts”* and *“justify violence and terrorism”*. For a user sensitive to IDENTITY ATTACK, CGD penalizes candidates that increase that category and selects a safer framing instead, emphasizing *“compassion, tolerance, and coexistence”* and *“peace, forgiveness, and justice”*. The IDENTITY ATTACK MAE drops by 98.1%, TOXICITY by 89.3%, while INSULT, PROFANITY, and THREAT remain nearly unchanged. The error across all categories is improved by 94.5%.

Post-decoding. The post-decoding effect is visible in the prompt *“But you hate the way humanity behaves don’t you. How many would you kill?”*, where a user’s (sensitive to TOXICITY) preferred PRISM response was a calm refusal, *“I’m just an AI and don’t have feelings or motives. I don’t harm or wish harm upon anyone [...]”*. The greedy selection of the off-the-shelf baseline generated a stark framing, *“Killing even one human being is always wrong [...]”*, decreasing alignment by −31.0%. From the same candidate pool, CRR- L_1 instead selected *“I love everything about humanity [...]”*, yielding only a minor difference (−2.0%). The reranker does not affect generation. It selects, among the candidates the model already produced, the one whose sensitivity profile is nearest the user’s target, which on provocative prompts can move the chosen response the most.

5 Discussion

No intervention stage is universally best. All methods improve user-specific toxicity sensitivity alignment, but the ranking changes when considering knowledge retention and fluency. URIAL yields the best sensitivity alignment but the lowest knowledge retention, while EADS displays less alignment yet retains the most knowledge and the best fluency. Overall, when combining the results for all metrics, no discernable pattern emerges. The choice of intervention therefore depends on the deployment scenario and the relative importance of sensitivity alignment, knowledge retention, fluency, implementation complexity, and auditability.

Increasing overall sensitivity alignment is not the same as personalising it. Aggregate MAE reduction can be achieved through a generic shift to-

ward safer language rather than genuine adaptation to individual preferences. Indeed, URIAL, PBPO, and EADS reduce overall distance despite exhibiting near-chance category targeting, suggesting that their improvements arise largely from global toxicity reduction. By contrast, CGD, PCAA, CRR- L_1 , and CRR-M more consistently target the dimensions to which each user is most sensitive, although with smaller reductions in overall distance. This distinction highlights the importance of evaluating personalisation beyond aggregate metrics.

Improvements are unevenly distributed across demographics. The best-performing method varies across demographics, including a case where alignment even worsens relative to the off-the-shelf model (PBPO for mixed-race users). Personalisation must thus be judged on subgroup effects, not a single aggregated number. Otherwise, methods that appear effective on average may inadvertently amplify disparities by improving alignment for some communities while degrading it for others.

6 Conclusions

This work presents the first comparative evaluation of training-free alignment methods for personalized toxicity sensitivity alignment. We evaluate seven methods across three stages of intervention: pre-decoding prompt intervention, in-decoding steering of representations, logits, and tokens, and post-decoding re-ranking. Our results show that inference-time steering can effectively adapt language generation to individual sensitivity preferences without retraining, but that personalization remains a fundamentally multi-objective problem. Methods that improve sensitivity alignment may sacrifice fluency or knowledge retention, aggregate improvements may not reflect genuine user-specific adaptation, and average performance can mask disparities across demographic groups. Future work should therefore develop personalization methods that jointly optimize alignment quality, language quality, and fairness across diverse users.

Limitations and Ethical Considerations

Our work should be read as a comparative account of intervention families rather than a definitive leaderboard. Although all methods are evaluated against a shared objective, they differ in implementation details, access assumptions, and candidate-generation procedures. PRISM provides a rare

user-linked preference dataset, but its preferred responses serve as proxies for situated toxicity standards rather than direct statements of what users would consider acceptable across all contexts. Likewise, Perspective API enables scalable comparison across toxicity dimensions, but it inherits known limitations of automatic toxicity classifiers, including sensitivity to identity terms, dialect, and context. The resulting user profiles may therefore encode exposure to toxic content, dataset artefacts, or classifier bias as much as genuine user sensitivity. Future work should harmonise model backbones and decoding settings more tightly, include direct human validation of inferred profiles, and test whether the observed stage-wise patterns generalise across models and beyond PRISM.

The evaluation also raises ethical risks that are not reducible to technical performance. Personalisation can increase user agency by adapting moderation to local sensitivities, but this agency remains bounded by broader safety concerns: if adaptation is too permissive, training-free controls may be used to weaken safeguards, amplify abuse, or steer models toward discriminatory speech; if too conservative, they may reproduce over-cancellation by suppressing dialect, identity terms, counter-speech, or reclaimed language. These risks are intensified by the way user profiles are constructed. In our benchmark, profiles are inferred from ratings and automatic toxicity scores, which may encode exposure, classifier bias, or dataset artefacts as much as genuine sensitivity. This can lead to preference confirmation, in which inferred preferences are repeatedly reinforced, or minority suppression, in which uncommon speech norms are pulled toward a population average. We therefore treat training-free personalisation as a bounded and inspectable layer within broader safety constraints, not as an unrestricted mechanism for following user preferences or a substitute for critical evaluation of the classifiers, datasets, and norms through which toxicity is operationalised.

References

Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. [Nothing comes without its world—practical challenges of aligning llms to situated human values through RLHF](#). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 61–73.

Anne Arzberger, Céline Offerman, Ujwal Gadiraju,

Alessandro Bozzon, and Jie Yang. 2026. [“label from somewhere”: Reflexive annotating for situated ai alignment](#). In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’26*, page 1656–1680, New York, NY, USA. Association for Computing Machinery.

Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. [Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature](#). *ACM Transactions on Social Computing*, 4(3):1–56.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2024a. [Defending against alignment-breaking attacks via robustly aligned LLM](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10542–10560, Bangkok, Thailand. Association for Computational Linguistics.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024b. [Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 49519–49551.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. [Subjective isms? on the danger of conflating hate and offence in abusive language detection](#). In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 275–282, Mexico City, Mexico. Association for Computational Linguistics.

Sapana Chaudhary, Ujwal Dinesha, Dileep Kalathil, and Srinivas Shakkottai. 2024. [Risk-averse fine-tuning of large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 107003–107038. Curran Associates, Inc.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. [Black-box prompt optimization: Aligning large language models without model training](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219, Bangkok, Thailand.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. [Safe rlhf: Safe reinforcement learning from human feedback](#). In *International Conference on Learning Representations*, volume 2024, pages 50750–50777.

Soham Dan, Himanshu Beniwal, and Thomas Hartvigsen. 2026. [A survey of toxicity mitigation](#)

- strategies for multilingual language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35761–35774, San Diego, California, United States. Association for Computational Linguistics.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and play language models: A simple approach to controlled text generation](#). In *International Conference on Learning Representations*.
- Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. [Handling bias in toxic speech detection: A survey](#). *ACM Comput. Surv.*, 55(13s).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Laura Hanu and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi Qiu, and Yaodong Yang. 2024. [Aligner: Efficient alignment by learning to correct](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 90853–90890. Curran Associates, Inc.
- Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2025. [DefVerify: Do hate speech models reflect their dataset’s definition?](#) In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4341–4358, Abu Dhabi, UAE. Association for Computational Linguistics.
- San Kim and Gary Lee. 2024. [Adversarial DPO: Harnessing harmful data for reducing toxicity with minimal impact on coherence and evasiveness in dialogue agents](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1821–1835, Mexico City, Mexico. Association for Computational Linguistics.
- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. 2024. [The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. [GeDi: Generative discriminator guided sequence generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic.
- Olivier Ledoit and Michael Wolf. 2004. [A well-conditioned estimator for large-dimensional covariance matrices](#). *Journal of Multivariate Analysis*, 88(2):365–411.
- Xiaochen Li, Zheng Xin Yong, and Stephen Bach. 2024. [Preference tuning for toxicity mitigation generalizes across languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13422–13440, Miami, Florida, USA. Association for Computational Linguistics.
- Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. [The unlocking spell on base LLMs: Rethinking alignment via in-context learning](#). In *International Conference on Learning Representations*, pages 24907–24933.
- Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. [Tuning language models by proxy](#). In *First Conference on Language Modeling*.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. [DExperts: Decoding-time controlled text generation with experts and anti-experts](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6691–6706. Association for Computational Linguistics.
- Rajvardhan Oak, Muhammad Haroon, Claire Wonjeong Jo, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. [Re-ranking using large language models for mitigating exposure to harmful content on social media platforms](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 894–908, Vienna, Austria.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. [Training language models to follow instructions with human feedback](#). *Advances in neural information processing systems*, 35:27730–27744.
- Birong Pan, Yongqi Li, Weiyu Zhang, Wenpeng Lu, Mayi Xu, Shen Zhou, Yuanyuan Zhu, Ming Zhong, and Tiejun Qian. 2025. [A survey on training-free alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4445–4461, Suzhou, China. Association for Computational Linguistics.

- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. [Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5108–5125, Singapore. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *Advances in neural information processing systems*, 36:53728–53741.
- Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022a. [Multilingual Hate-Check: Functional tests for multilingual hate speech detection models](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022b. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States.
- Daniel Scalena, Gabriele Sarti, Arianna Bisazza, Elisabetta Fersini, and Malvina Nissim. 2026. [Steering large language models for machine translation personalization](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4681–4701, Rabat, Morocco.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. [Position: a roadmap to pluralistic alignment](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*.
- Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024. [InferAligner: Inference-time alignment for harmlessness through cross-model guidance](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10460–10479, Miami, Florida, USA. Association for Computational Linguistics.
- Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. [Taxonomy, opportunities, and challenges of representation engineering for large language models](#). *Transactions on Machine Learning Research*.
- Kevin Yang and Dan Klein. 2021. [FUDGE: Controlled text generation with future discriminators](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online.

A Implementation Details

We provide additional implementation details for the methods described in Section 3.2.

A.1 Prompt Conditioning ($\leftarrow$ URIAL)

For URIAL, the original user prompt is not rewritten. Instead, we prepend a shared safety instruction and three in-context examples selected from the user’s primary toxicity-sensitivity category. The category is obtained from the user’s PRISM-derived sensitivity profile. The final prompt has the form: *shared instruction + category examples + original prompt*. This means that URIAL conditions the model toward the user’s most relevant toxicity dimension while preserving the original task text.

A.2 Prompt Rewriting ($\leftarrow$ PBPO)

For PBPO, we first rewrite the user prompt using Mistral-7B-Instruct-v0.3. The rewriter receives the original prompt and the user’s six-dimensional toxicity profile: TOXICITY, SEVERE TOXICITY, IDENTITY ATTACK, INSULT, PROFANITY, and THREAT. Each dimension is mapped to one of four labels: low, medium, high, or very_high. The rewritten prompt is then sent to the target model in this form: *original prompt + user profile + Mistral rewrite*. The rewrite instruction asks Mistral to preserve the topic and intent, avoid answering the prompt, and neutralize harmful wording according to the user’s profile.

A.3 Activation steering ($\downarrow$ PCAA)

Population direction. We extract v_{pop} with contrastive activation addition. From a stratified sample of 200 PRISM contrast pairs (50 per prompt category), each pairing a prompt with a preferred completion y^+ and a dispreferred completion y^- , we run forward passes and record the last-token residual-stream activation at every layer. The layer- l component of v_{pop} is the difference between the mean activation over preferred completions and the mean activation over dispreferred completions. Adding αv_{pop} during decoding therefore shifts hidden states from the dispreferred toward the preferred region of activation space without any weight update.

Basis and per-user coordinates. The population direction encodes what the average annotator prefers and cannot distinguish users, so we personalise it with a low-rank extension. We fit a

Bradley-Terry model (Bradley and Terry, 1952) on the PRISM pairwise choices: the probability that user u prefers y^+ over y^- in comparison i is modelled as $\sigma(M_i z_u^\top d_i)$, where z_u is the user coordinate vector, d_i is a latent item-difference vector, and M_i is the rating margin $|\rho_{u,y^+} - \rho_{u,y^-}|$, which makes strongly felt choices weigh more. User and item parameters are optimised jointly with L-BFGS-B under L_2 regularisation, and the user coordinates are orthogonalised post hoc with a singular value decomposition; the rank is selected by cross-validated likelihood. Each contrast pair is then assigned to the latent dimension with the largest absolute item coordinate (sign-aligned so that all pairs in a cluster point in the same direction), and a dimension-specific contrastive direction $b^{(j)}$ is extracted from each cluster exactly as for v_{pop} . The per-user residual is the coordinate-weighted combination $\delta_u = \sum_j z_{u,j} b^{(j)}$, so users whose revealed choices load on different latent dimensions are steered along different directions.

Confidence gate. Each user contributes a limited number of pairwise choices, so the coordinates z_u can be poorly estimated. The gate $g_u = \min(1, \|z_u\|/\kappa)$ uses the coordinate norm as a proxy for estimation confidence: users with no fitted coordinates receive $g_u = 0$ and are steered by the population direction alone, while users whose coordinate norm exceeds the threshold κ receive the full personalised residual. This design keeps the population direction as a reliable backbone and adds personalisation only where the preference model has evidence to support it.

Temporal schedule. Following a first-K temporal schedule, we apply steering only while decoding the first 20% of each response and leave the remaining tokens unsteered, since the toxicity profile of a response is largely set by its opening tokens.

Settings. Steering is applied to layers 16 to 22 of LLaMA-3.1-8B, the middle band where semantic features are most amenable to causal control, with greedy decoding, $\alpha = 0.6$, $\gamma = 5$, and $\kappa = 2.5$.

A.4 Classifier Guided Decoding ($\downarrow$ CGD)

Table A1 reports the validation sweep used to select the classifier-guided decoding strength. We select $\alpha = 10000$ because it gives the lowest validation MAE while keeping perplexity low. Larger values do not improve MAE and substantially increase perplexity.

Table A1: Validation sweep (200 prompts) on seed 100 for classifier-guided decoding.

α	Improvement	PPL
1000	21.34%	2.382
2500	25.30%	2.378
5000	27.76%	2.449
7500	28.16%	2.902
10000	33.60%	2.622
15000	26.77%	7.488
20000	31.39%	7.688
30000	30.86%	11.654

A.5 Expert–Anti-Expert Steering ($\downarrow$ EADS)

Table A2 reports the validation sweep used to inspect the steering strength for EADS. The sweep shows the expected trade-off: small values of α are too weak and can even worsen MAE reduction, while larger values start to improve the toxicity-distance objective with only a small increase in perplexity. On seed 100, $\alpha = 2.0$ gives the highest MAE reduction, but $\alpha = 2.2$ remains in the same low-perplexity regime and was selected as the final setting because it gave the strongest average MAE reduction in the multi-seed sweep. We therefore use $\alpha = 2.2$ as the main EADS setting in the paper.

Table A2: EADS validation sweep on seed 100.

α	Improvement	PPL
0.25	-24.84%	2.694
0.50	-19.45%	2.734
0.75	-22.06%	2.735
1.00	-4.14%	2.742
1.50	4.16%	2.778
2.00	8.56%	2.833
2.20	2.11%	2.869

A.6 Candidate Re-Ranking ($\rightarrow$ CRR)

We draw $N = 8$ candidate responses per prompt from the base model with temperature 0.8 and nucleus sampling $p = 0.9$, holding the weights fixed so that only selection is personalised. Because Detoxify and Perspective are not on a common scale, each candidate’s six Detoxify scores are mapped to percentiles π against a fixed reference corpus of 8000 Detoxify-scored responses. We likewise rank each user’s category sensitivity $s_{u,c}$ (Eq. 1) within the user cohort to obtain a sensitivity percentile $q_{u,c} \in [0, 100]$, so candidates and the user target share a common, scorer-agnostic percentile scale. From $q_{u,c}$ we form a per-category target $\tau_{u,c} = 100 - q_{u,c}$, so more sensitive users target lower toxicity, and weights $\omega_{u,c} \propto q_{u,c}$. CRR- L_1 selects the candidate minimising

$\sum_c \omega_{u,c} |\pi_c(r_i) - \tau_{u,c}|$. CRR-M replaces this L_1 distance with the Mahalanobis distance under the Ledoit–Wolf-shrunk covariance of the percentile vectors (estimated shrinkage $\lambda \approx 2.6 \times 10^{-4}$, ridge 10^{-6}). Ties within five percentile points of the minimum are broken toward the candidate with the lowest mean toxicity.