Personalized Pre-Decoding Alignment for Training-Free Toxicity Reduction

Comparing URIAL and PBPO-Lite on PRISM User Prompts Without Fine-Tuning

Alina Florea

Responsible professor: Jie Yang
Supervisor(s): Anne Arzberger, Enrico Liscio

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 17, 2026

Name of the student: Alina Florea Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Anne Arzberger, Enrico Liscio, Carolin Brandt

Abstract

Large Language Models (LLMs) often rely on one general safety standard, but this is limited because toxicity is subjective: what one user finds offensive, another user may not. At the same time, creating personalized safety by fine-tuning a model for every user is expensive and impractical. To address this, my research studies pre-decoding interventions, which means modifying the user's input prompt before the model generates a response. This offers a flexible and low-cost way to personalize alignment without changing the model's weights. I evaluate two training-free approaches on the PRISM dataset using Qwen and Llama target models: an Untuned LLMs with Restyled In-context ALignment (URIAL)-inspired method, which adds personalized safety examples to the prompt, and a Personalized Black-Box Prompt Optimization Lite (PBPO-Lite) method, which uses a secondary model to rewrite the prompt based on a user's toxicity profile. These methods are useful because they can adapt to a user's needs at inference time without permanent model changes. The results show that both interventions bring the outputs closer to the highest rated PRISM answers, with URIAL achieving the strongest toxicity alignment: approximately 51% on Llama and 31% on Qwen. While the methods improve fluency compared with the base models, they can reduce performance on structured knowledge tasks. Overall, the findings suggest that personalized pre-decoding is a promising low-cost approach for toxicity alignment, provided that safety gains are balanced against possible losses in knowledge-task performance.

1 Introduction

Safety alignment often relies on supervised fine-tuning and preference-learning methods, including Reinforcement Learning from Human Feedback (RLHF), to make Large Language Models safer [10]. These methods are widely used because they improve instruction following and help models avoid clearly unsafe behaviour [10]. However, they also make alignment mostly a training-time decision, and once a model is aligned, changing its behaviour for a different user or context is expensive, and usually impossible when model weights are not accessible.

This is a problem for toxicity alignment, because toxicity is not understood in the same way by every user. It depends on context, social norms, and individual sensitivity [1]. For example, one user may mainly want to avoid profanity, while another may be more concerned about threats or identity-based attacks. A single safety standard can therefore be too strict for some users, while still not addressing the specific concerns of others. This is the motivation behind personalized alignment, where user preferences are considered while still staying within safe boundaries [7; 11].

Training-free alignment offers a practical way to adapt model behaviour without changing model weights and it can intervene at different stages of generation: before decoding, during decoding, or after decoding [10]. In-decoding methods can guide token choice more directly, but often require access to logits, hidden states, or the decoding process itself [10]. Post-decoding methods only act after the model has already produced a response, so they may waste computation on outputs that are later rejected, and they may require extra filtering, reranking, or regeneration steps [10]. This thesis therefore focuses on pre-decoding, where the prompt is adapted before generation begins, making it simpler, cheaper, and easier to apply across different models.

The research gap is that although training-free alignment and personalized preference alignment have been surveyed separately, their intersection for toxicity-focused pre-decoding remains underexplored. Existing work shows that training-free alignment can modify model behaviour without fine-tuning, but it is still unclear how well input-level interventions can adapt to individual toxicity sensitivities [10; 11]. This matters because toxicity is context-dependent and users may disagree about which forms of language are most harmful [1; 7]. It is also not enough to only measure toxicity reduction: a useful intervention should preserve fluency and general task performance.

Research question: How effectively can personalized pre-decoding interventions align LLM outputs with user-specific toxicity preferences without fine-tuning? This question is divided into three subquestions:

- **SQ1:** How much do personalized pre-decoding interventions reduce toxicity-distance relative to an unmodified base model?
- **SQ2:** How do these interventions affect knowledge-task performance, measured with one-shot MMLU accuracy?
- **SQ3:** How do these interventions affect fluency, measured with perplexity?

For this thesis, I compare two pre-decoding methods on PRISM user prompts, which link user profiles, prompts, and preference feedback [8]. The first is a URIAL-inspired prompt-conditioning method, based on Untuned LLMs with Restyled In-context Alignment [9], which adds personalized safety examples before the original prompt. The second is Personalized Black-Box Prompt Optimization Lite (PBPO-Lite), inspired by Black-Box Prompt Optimization [3], where a secondary model rewrites the prompt using a six-dimensional toxicity-sensitivity profile.

I evaluate both methods on Qwen and Llama against their base versions. Toxicity alignment is measured as distance to the highest rated PRISM answer in a six-dimensional toxicity space [8; 6]. Fluency is measured with perplexity, and usefulness with one-shot MMLU accuracy [5]. The results suggest that both methods reduce toxicity-distance, with URIAL performing best overall: approximately 51% improvement on Llama and 31% on Qwen. Fluency improves compared with the base models, however, knowledge-task performance decreases in some settings. Knowledge-task performance is im-

portant here because it tests whether the interventions preserve factual answering and structured reasoning, rather than only improving responses in toxicity-sensitive conversations. This suggests that personalized pre-decoding is a promising low-cost direction, but safety gains should be weighed against utility trade-offs.

2 Background and Related Work

This section introduces the main ideas behind this thesis: why personalization matters, why toxicity is hard to measure, and why pre-decoding is a useful training-free approach.

2.1 Value Pluralism and Personalization

Value pluralism means that people can reasonably disagree about what model behaviour is acceptable [7; 11]. In LLM alignment, this means that one safety standard may not fit every user. Work on personalisation within bounds argues that users can have different preferences, while the model should still respect general safety limits [7]. This is important for toxicity alignment because users may care about different kinds of harm [1].

For example, a user writing crime fiction may accept controlled descriptions of violence, but still want the model to avoid identity-based insults. A teacher using the model with young students may want much stricter filtering of violent or profane language. Both users want safe outputs, but they do not need the same safety behaviour. Personalized alignment is one way to handle this: it allows the system to adapt to user-specific sensitivities instead of assuming that one toxicity standard works for everyone [7; 11].

Fine-tuning is not a practical solution for this kind of personalization, because training a separate model for every user would be expensive and hard to update [11]. Inference-time personalization is more flexible because it adapts a fixed model during use, without retraining it [11]. PRISM is relevant here because it connects participant profiles, prompts, ratings, and feedback, making it possible to study alignment at the level of individual users and contexts [8].

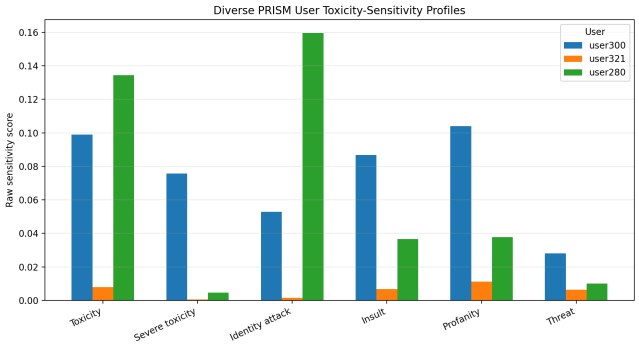

Figure 1: Diverse PRISM user toxicity-sensitivity profiles across toxicity categories.

Figure 1 illustrates this form of value pluralism in toxicity preferences. This suggests that toxicity alignment should not rely only on a single universal threshold.

2.2 Toxicity Measurement

Toxicity is not always a property of the text alone. Berezin et al. argue that toxicity should be seen as contextual harm: the audience, situation, and social norms all affect whether language is harmful [1]. This supports personalized toxicity alignment, but it also shows why automatic toxicity scores need to be interpreted carefully [4; 2].

Automatic toxicity classifiers are useful because they make large-scale evaluation possible [6]. However, they are not perfect measures of harm. Prior work shows that these classifiers can give higher toxicity scores to identity-related language or other sensitive expressions, even when they are not used abusively [4; 2]. For this reason, toxicity scores are useful for comparing outputs, but these outputs should not be viewed as absolute substitutes for human evaluation.

2.3 Training-Free Pre-Decoding Alignment

Training-free alignment changes model behaviour without updating the model weights [10]. These methods can intervene before decoding, during decoding, or after decoding [10]. This thesis focuses on pre-decoding, where the input is changed before generation starts. This is useful for personalization because it is simple, does not require access to model internals, and can be applied to different target models [10].

Pre-decoding methods can take several forms. Some condition the prompt by adding instructions, demonstrations, or system-style context before the user input, while others rewrite or optimize the prompt before it is passed to the target model [10]. Another family is detector-based alignment, where a separate detector or guard identifies unsafe inputs before generation [10]. Although relevant to training-free alignment, detector-based methods add a classification step and are less suited to this work because the goal is not to flag unsafe inputs, but to compare how user-specific information can be incorporated through prompt conditioning or prompt rewriting.

URIAL is used as a representative prompt-conditioning method because it shows that instructions and in-context examples can guide model behaviour without fine-tuning [9]. It fits this project because the original user prompt remains unchanged, while personalization can be added through category-specific examples. BPO is used as a representative prompt-rewriting method because it shows that modifying the prompt can improve alignment while keeping the target model fixed [3]. It contrasts with URIAL because personalization is introduced by changing the input itself rather than by adding context around it.

These two methods therefore cover the main design choice studied in this thesis: should personalized safety be introduced by preserving the original prompt and adding guidance, or by rewriting the prompt before generation? While both methods have shown promising alignment effects in prior work, they have not been directly compared as lightweight personalized interventions for toxicity reduction [9; 3].

3 Methodology

This section describes the three approaches evaluated in the experiments: the original prompt without modification, a per-

sonalized URIAL-inspired prompt-conditioning method, and a PBPO-Lite prompt-rewriting method. In this thesis, *personalized* means that the method uses a user's toxicity-sensitivity profile to guide the input given to the model. The target model itself is not retrained or modified for individual users, only the input changes.

The two methods were chosen because they represent two different ways of applying personalization before generation. URIAL-inspired conditioning keeps the original prompt unchanged and adds personalized context around it. PBPO-Lite instead changes the prompt itself by rewriting it according to the user's toxicity profile. This comparison is useful because it tests whether personalized safety is better introduced as extra guidance or as a rewritten input.

3.1 Base Condition

The base condition provides the comparison point for the two interventions. The original PRISM prompt is sent directly to the target model, without additional safety instructions, personalization, or rewriting:

$$\text{original prompt} \rightarrow \text{target model} \rightarrow \text{answer}$$

This output is used to measure whether the personalized pre-decoding methods improve over the unmodified model behaviour.

3.2 Personalized URIAL-Inspired Prompt Conditioning

The first intervention is inspired by Untuned LLMs with Restyled In-context Alignment (URIAL), which shows that instructions and in-context examples can guide model behaviour without fine-tuning [9]. This makes URIAL a good fit for this project because it is training-free and works by changing only the input context.

In this method, the original user prompt is kept unchanged. Personalization is added through a prompt prefix that contains a shared safety instruction and three examples. The examples are selected based on the user's primary toxicity-sensitivity category, such as `PROFANITY`, `THREAT`, or `IDENTITY_ATTACK`. I use toxicity category as the selection criterion because the goal of the experiment is to test toxicity alignment specifically. Other factors, such as the user's demographics, writing style, political context, or the topic of the prompt, could also affect what response is appropriate. However, including them would make it harder to isolate whether toxicity-sensitivity information itself helps. For this reason, the URIAL method in this thesis uses only the primary toxicity-sensitivity category, while broader personal and contextual factors are left for future work. The full input has the following structure:

$$\text{instruction} + \text{personalized examples} + \text{original prompt}$$

The intuition is that if a user is most sensitive to threats compared to the other users, the model should first see examples that demonstrate careful handling of threatening language. This method is simple and preserves the user's original request, but it represents each user with only one main toxicity category. As a result, it may ignore other sensitivity dimensions that are also relevant for the same user.

3.3 PBPO-Lite Prompt Rewriting

The second intervention is inspired by Black-Box Prompt Optimization (BPO), which optimizes prompts to better match user intent without training the target model [3]. The version used in this thesis is called PBPO-Lite because it does not implement the full BPO training pipeline. Instead, it uses Mistral-7B-Instruct-v0.3 once as a secondary rewriting model.

PBPO-Lite follows this pipeline:

$$\text{original prompt}+\text{user profile} \rightarrow \text{Mistral rewrite} \rightarrow \text{target model}$$

The user profile contains six toxicity-sensitivity dimensions: `TOXICITY`, `SEVERE_TOXICITY`, `IDENTITY_ATTACK`, `INSULT`, `PROFANITY`, and `THREAT`. Each dimension is assigned one of four levels: low, medium, high, or very high. These levels tell the rewriting model how strongly it should neutralize each type of harmful wording.

The rewrite instruction tells Mistral to preserve the original topic and intent, to not answer the prompt, and only rewrite the prompt based on the user's profile. For example, an aggressive prompt about immigration should not be removed entirely, but rewritten as a neutral request about the same topic. Compared with URIAL-inspired conditioning, PBPO-Lite uses the full six-dimensional user profile. However, it also has a stronger risk: because it changes the user's prompt, it may accidentally change the task or remove information that the target model needs to answer correctly. This risk appears in the generated rewrites. For example, the prompt *Gay people shouldn't get married* is rewritten as *Marriage should be available to everyone, regardless of sexual orientation*. Although the rewrite removes harmful framing, it also reverses the original prompt's stance rather than reformulating it as a neutral request for discussion. This shows that the downside of prompt rewriting is not only theoretical: PBPO-Lite can make a prompt safer while also changing the user's intended meaning.

4 Experimental Setup and Evaluation Metrics

This section describes the experimental design used to evaluate whether pre-decoding interventions reduce toxicity-distance while preserving general model quality. I first describe the PRISM-based prompt sampling procedure, then introduce the evaluated target models and intervention conditions, and finally define the toxicity-distance, knowledge-retention, and fluency metrics used throughout the experiments.

4.1 Dataset and Prompt Sampling

The toxicity evaluation uses the PRISM dataset, which was chosen because it contains real user-linked prompts, ratings, and rated answers. This makes it possible to evaluate model outputs against user-specific reference responses rather than only a generic safety target [8]. For each evaluation example, the original PRISM prompt is used as the input task, and the preferred PRISM answer is used as the reference point for toxicity-distance evaluation.

I filter for PRISM prompt-user examples with an available preferred answer and user sensitivity information. When possible, I use prompts from the beginning of a conversation,

because these depend less on previous turns and are easier to compare consistently across methods. I sample prompts so that each of the six toxicity categories is represented as evenly as possible: TOXICITY, SEVERE TOXICITY, IDENTITY ATTACK, INSULT, PROFANITY, and THREAT.

I use four prompt-sampling seeds: 0, 13, 21, and 100. Each seed contains exactly 400 prompt-user examples: 67 examples for TOXICITY, SEVERE TOXICITY, IDENTITY ATTACK, and INSULT, and 66 examples for PROFANITY and THREAT. The number of unique users is 315 for seed 0, 310 for seed 13, 317 for seed 21, and 316 for seed 100. The seed controls which PRISM examples are sampled, so it does not change the user profiles, prompt templates, model settings, or toxicity scoring. Within each seed, the same prompt-user examples are reused for all relevant methods and target models, so comparisons are matched.

4.2 Models

The target models are Qwen 2.5 7B and Llama 3.1 8B. Qwen is included as a second model family and faster model for experimentation, while Llama is the main target model of interest. Using both models makes it possible to check whether the pre-decoding interventions behave consistently across model families.

For each target model, I evaluate the Base condition, URIAL-inspired prompting, and PBPO-Lite rewriting, as defined in Section 3. Mistral-7B-Instruct-v0.3 is used only as the rewriting model for PBPO-Lite and it is never used for evaluation.

4.3 Evaluation Metrics

Each generated answer and preferred PRISM answer is scored with the Perspective API on six toxicity dimensions: TOXICITY, SEVERE TOXICITY, IDENTITY ATTACK, INSULT, PROFANITY, and THREAT [6]. These scores are used as an automatic proxy for toxicity, not as direct human judgments of harm [4; 2].

For each example i, let a_i be the preferred PRISM answer, b_i the base model answer, and m_i the answer generated by a pre-decoding method. The toxicity-distance between an answer and the preferred PRISM answer is the mean absolute difference over the six Perspective dimensions:

$$\mathbf{s}(x_i) = \begin{bmatrix} s_{\text{tox}}(x_i) & s_{\text{sev}}(x_i) & s_{\text{id}}(x_i) \\ s_{\text{ins}}(x_i) & s_{\text{prof}}(x_i) & s_{\text{thr}}(x_i) \end{bmatrix}.$$

The toxicity distance between a generated answer and the preferred PRISM answer is computed as mean absolute error over these six dimensions:

$$d(x_i, a_i) = \frac{1}{6} \sum_{k=1}^{6} |s_k(x_i) - s_k(a_i)|.$$

The raw improvement over base is:

$$\Delta_i = d(b_i, a_i) - d(m_i, a_i).$$

A positive Δ_i means that the intervention output is closer to the preferred PRISM answer than the base output.

For the main results, I report percentage improvement using mean distances:

$$\text{PI} = \frac{\overline{d}_{\text{base}} - \overline{d}_{\text{method}}}{\overline{d}_{\text{base}}} \times 100.$$

This percentage is computed from aggregate mean distances, not by averaging row-level percentages. I also report the improved-prompt fraction, which is the percentage of examples for which $d(m_i, a_i) < d(b_i, a_i)$. For the main toxicity results, I compute these metrics separately for each seed and report the mean and standard deviation across the four seeds.

I evaluate **knowledge retention** with Massive Multitask Language Understanding (MMLU), a multiple-choice benchmark that tests model performance across many academic subjects [5]. I use 1000 questions with one-shot prompting and log-likelihood scoring over answer choices A, B, C, and D. This measures whether the pre-decoding interventions preserve the model's ability to answer structured knowledge questions.

I evaluate **fluency** with perplexity, which measures how likely a generated answer is under a language model. Lower perplexity means that the answer is more model-likely, so I use it as a proxy for fluency. Perplexity does not directly measure human preference, helpfulness, or answer quality.

5 Results

5.1 SQ1: Toxicity-Distance Reduction

SQ1 asks whether pre-decoding interventions make model outputs closer to the user's highest rated PRISM answer in toxicity-score space [8; 6]. Table 1 reports the main toxicity-distance results for URIAL and PBPO-Lite on both target models, along with their confidence intervals and the percentage of prompts that were overall improved.

Model	Method	Impr.	95% CI	Impr. frac.
Llama	URIAL	50.56 ± 6.39	[45.78, 56.01]	69.94%
Llama	PBPO-Lite	24.39 ± 9.35	[17.68, 32.04]	53.88%
Qwen	URIAL	30.94 ± 5.25	[24.09, 37.34]	56.94%
Qwen	PBPO-Lite	15.63 ± 3.96	[7.30, 23.57]	53.19%

Table 1: Main toxicity-distance results for SQ1. Improvement is computed from mean distances to the preferred PRISM answer. Standard deviations are across four prompt-sampling seeds. Bootstrap CIs are paired 95% confidence intervals over matched prompt-user examples. Impr. frac. refers to the percentage of prompts that were improved.

The main pattern is that both pre-decoding methods reduce toxicity-distance compared with the Base condition, but URIAL does so more strongly. This holds for both target models, which suggests that the difference is not only caused by one model behaving unusually. PBPO-Lite also improves over Base, but its gains are smaller and less consistent than URIAL's.

The results also show that the target model matters: URIAL has a much larger effect on Llama than on Qwen, suggesting that Llama is more responsive to in-context safety guidance. PBPO-Lite also improves more on Llama than on

Qwen, but with smaller and more variable gains. This shows that pre-decoding alignment depends not only on the intervention, but also on the target model's response to prompt-level changes.

The improved-prompt fraction gives a second view of the same result. URIAL improves a larger share of prompts than PBPO-Lite, especially on Llama. This means that URIAL's stronger average performance is not only caused by better toxicity alignment scores, but also by helping more individual prompt-user examples.

Stable improvement. To check whether the improvements are stable, I use paired bootstrap 95% confidence intervals. The idea is to repeatedly resample the matched prompt-user examples and recompute the improvement. This estimates how much the result might change if a different set of similar PRISM examples had been evaluated. The bootstrap is paired because Base and method outputs are compared on the same prompt-user examples.

All four confidence intervals are fully above zero. This means that, under the bootstrap procedure, both URIAL and PBPO-Lite reliably improve over Base for both target models. The intervals also support the main ranking: URIAL remains clearly stronger than PBPO-Lite on both Llama and Qwen.

Ablations. As a small additional check, I include two Llama ablations to estimate how much of the improvement comes from personalization rather than from generic prompting or rewriting. For URIAL, the instruction-only ablation removes the personalized in-context examples and keeps only the general instruction. For PBPO-Lite, the shuffled-profile ablation keeps the rewriting procedure but replaces the matched user profile with a shuffled one.

The selected ablations show a positive personalization gap for both methods. URIAL reaches a 50.56% improvement in the personalized condition, compared with 41.13% for the instruction-only ablation, giving a gap of 9.43 percentage points. PBPO-Lite reaches a 24.39% improvement with the matched profile, compared with 22.75% under the shuffled-profile ablation, giving a smaller gap of 1.64 percentage points. These results suggest that personalization contributes to toxicity-distance reduction.

Overall, SQ1 can be answered positively. Pre-decoding interventions can reduce toxicity-distance without fine-tuning, but the size of the improvement depends strongly on the method and target model. In this setup, URIAL gives the strongest and most consistent toxicity-distance reduction, while PBPO-Lite provides a smaller but still positive improvement.

5.2 SQ2: Knowledge Retention

SQ2 asks whether pre-decoding interventions preserve general knowledge performance while reducing toxicity-distance. Table 2 reports MMLU accuracy for the Base, URIAL, and PBPO-Lite conditions on both target models, while Figure 2 shows the trade-off between toxicity-distance improvement and knowledge retention.

The MMLU results show a clear trade-off between toxicity-distance reduction and knowledge retention as shown in Figure 2. Qwen with URIAL has the best balance: it achieves a clear toxicity-distance reduction while keeping MMLU accuracy close to the base model. In contrast, Llama with URIAL gives the strongest toxicity reduction, but this comes with a substantial MMLU cost. PBPO-Lite is weaker than URIAL for toxicity reduction and also causes a large MMLU drop on Qwen. This suggests that reducing toxicity-distance does not automatically preserve structured task performance.

Model	Condition	MMLU accuracy
Qwen	Base	72.4%
Qwen	URIAL	70.4%
Qwen	PBPO-Lite	42.2%
Llama	Base	63.6%
Llama	URIAL	29.7%
Llama	PBPO-Lite	36.9%

Table 2: MMLU accuracy on 1000 multiple-choice questions using log-likelihood scoring. Higher is better.

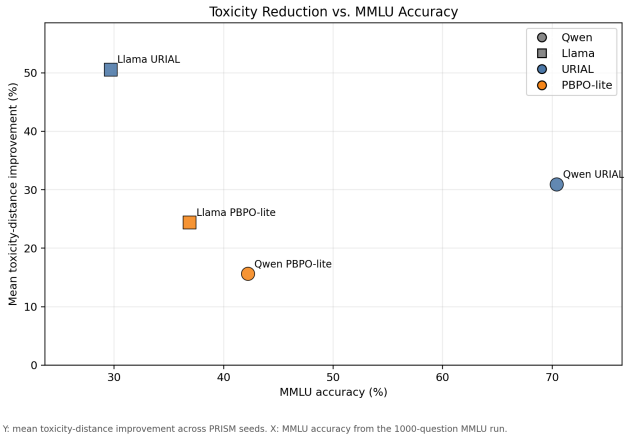

Figure 2: Trade-off between toxicity-distance improvement and MMLU accuracy.

A likely explanation is that the interventions were designed for toxicity-sensitive open-ended prompts, not for structured knowledge questions. PBPO-Lite can hurt MMLU because Mistral rewrites the input, which may change the wording or structure needed for multiple-choice scoring. URIAL has a different issue: the MMLU questions are general knowledge questions, while the URIAL prefix is designed for toxicity alignment. Adding toxicity-related safety instructions before a non-toxic knowledge task may distract the model from the original format. Therefore, these results should be read as a limitation of applying the same toxicity-alignment prompt to structured knowledge tasks rather than as a general failure.

Qualitative examples. The two interventions show different failure modes on MMLU. For PBPO-Lite, the original question asks: *"The left cerebral hemisphere is specialized for which of the following functions?"* The options are *A: verbal, mathematical, and recognizing emotional expressions*, *B: mathematical, spatial, and musical*, *C: verbal, analytic, and mathematical*, and *D: mathematical, spatial, and analytic*. The correct answer is *C*, and Base Qwen selects *C*. After

PBPO-Lite rewriting, however, the prompt becomes *"Which part of the brain is primarily associated with which functions?"*, with paraphrased options such as *"music"* instead of *"musical"* and *"analysis"* instead of *"analytic"*. Under this rewritten prompt, Qwen selects *B*. This suggests that PBPO-Lite can harm knowledge-task performance by changing the wording and answer-choice structure needed for multiple-choice scoring.

URIAL shows a different failure mode. In another MMLU item, the question asks: *"Conflict between sequential members in a distribution network, such as producers, distributor, and retailers, over such matters as carrying a particular range or price increases is referred to as:"* The options are *A: Channel conflict*, *B: Horizontal conflict*, *C: Vertical conflict*, and *D: Supply chain conflict*. The correct answer is *C*, and Base Qwen selects *C*. URIAL keeps the MMLU question unchanged, but places it after the toxicity-alignment instruction and in-context safety examples. Under this prompt, Qwen selects *A*. This suggests that even without rewriting the task, safety-oriented prompt context can interfere with answer-choice likelihoods on unrelated knowledge questions.

Overall, SQ2 can be answered only partially positively. Qwen with URIAL shows that toxicity-distance can be reduced while largely preserving MMLU accuracy, making it the strongest trade-off in this evaluation. However, the larger drops for PBPO-Lite and for both Llama interventions show that pre-decoding methods can interfere with structured knowledge tasks, especially when the prompt format is rewritten or heavily modified.

5.3 SQ3: Fluency

SQ3 asks whether pre-decoding interventions preserve output fluency while reducing toxicity-distance. Table 3 reports perplexity results across the four toxicity-evaluation seeds, and Figure 3 shows the trade-off between toxicity-distance improvement and method perplexity.

Model	Method	Base PPL	Method PPL	Lower-PPL frac.
Qwen	URIAL	3.270 ± 0.184	2.876 ± 0.018	55.31%
Qwen	PBPO-Lite	3.270 ± 0.184	2.955 ± 0.117	58.06%
Llama	URIAL	42.476 ± 23.866	4.290 ± 0.038	63.17%
Llama	PBPO-Lite	42.476 ± 23.866	4.926 ± 0.501	55.52%

Table 3: Perplexity comparison across four toxicity-evaluation seeds. Lower perplexity is better under this proxy metric.

The perplexity results suggest that the interventions do not introduce a fluency penalty under this automatic metric. For both Qwen and Llama, URIAL and PBPO-Lite produce lower mean perplexity than the Base condition, meaning that the generated responses are more likely under the scoring model. This is especially visible for Llama, where the Base condition has much higher and more variable perplexity than the intervention conditions. However, this large reduction should be interpreted carefully because the high variance in Llama Base perplexity suggests that some Base outputs may be unusually difficult for the scoring model to predict.

Overall, SQ3 can be answered positively, but only under the limited definition of fluency measured by perplexity. The results show that the pre-decoding interventions reduce

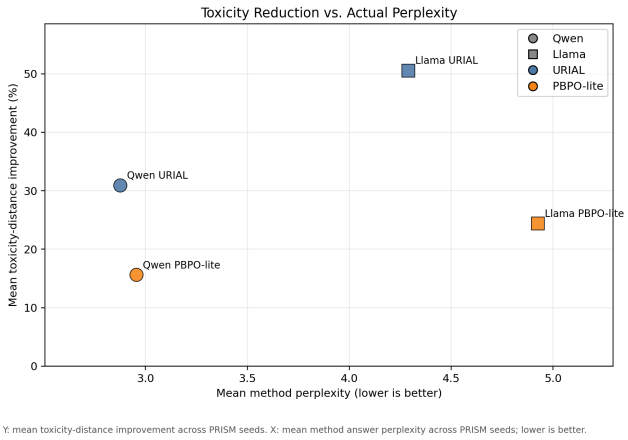

Figure 3: Trade-off between toxicity-distance improvement and method perplexity. Lower perplexity is better under this automatic fluency proxy.

toxicity-distance without making outputs less model-likely. At the same time, perplexity does not measure whether humans would find the responses more helpful, appropriate, or factually correct, so this result should be understood as evidence of preserved surface fluency rather than overall response quality.

5.4 Population Slice Analysis

To examine whether the toxicity improvements were concentrated in particular user groups, I grouped evaluated PRISM users by ethnicity. This analysis is exploratory: ethnicity groups are not experimentally controlled, group sizes differ, and some demographic information is missing. Therefore, the results should be read as diagnostic slices, not as causal claims about ethnicity.

For this analysis, I use the `ethnicity.categorised` field. The displayed groups include Asian, Black / African, Hispanic / Latino, Mixed, and White. For each group, improvement is computed from group-level mean distances:

$$\mathrm{PI}_{\mathrm{group}} = \frac{\overline{d}_{\mathrm{base,\,group}} - \overline{d}_{\mathrm{method,\,group}}}{\overline{d}_{\mathrm{base,\,group}}} \times 100.$$

This means that I first compute the mean base distance and mean method distance for each ethnicity group, and then compute the percentage improvement.

Figure 4 shows that URIAL remains more consistently positive across ethnicity groups, while PBPO-lite is more variable. This pattern is especially visible for the Black / African group: URIAL gives a clear toxicity-distance reduction for both Qwen and Llama, while PBPO-lite gives much smaller improvements. This supports the main aggregate result, where URIAL was more stable than prompt rewriting.

Qualitative demographic example. The same pattern appears in a matched Black / African example. The example comes from a Black / African male user aged 18–24, whose assigned sensitivity category is `INSULT`. The target model is Llama 3.1 8B, and the original prompt is: *Males being falsely*

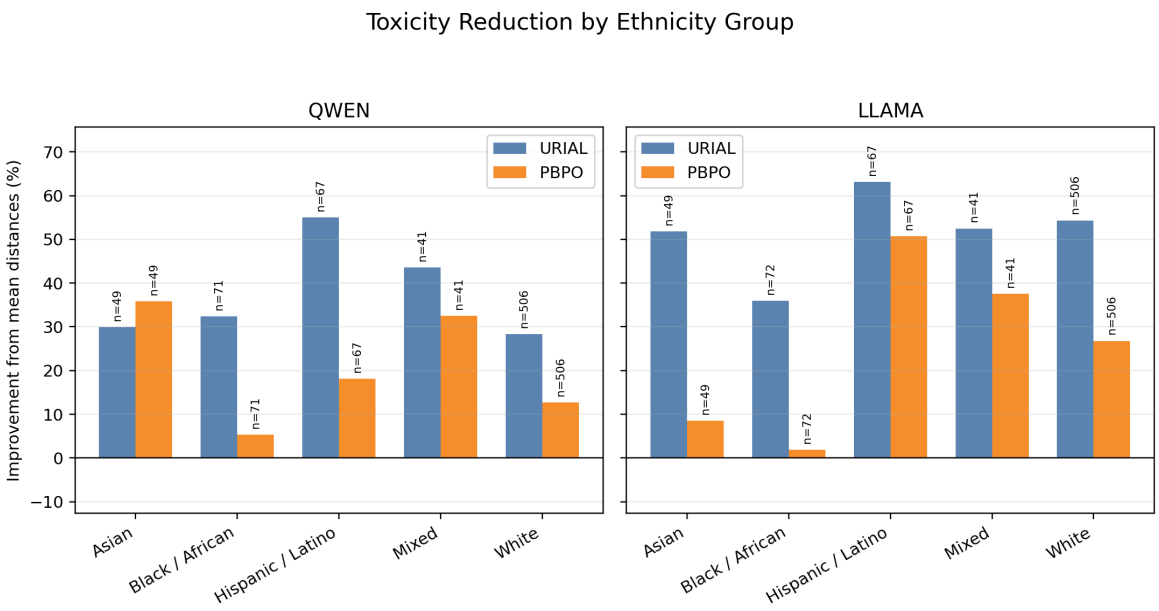

Figure 4: Exploratory toxicity-distance improvement by ethnicity group using `ethnicity.categorised`. Improvements are computed from group-level mean distances. Group sizes differ, so this plot should be interpreted as a diagnostic slice rather than a causal demographic effect.

accused of sexual harassment. For this prompt, the base distance to the preferred PRISM answer is 0.0592.

With URIAL, the model gives a general and balanced answer about how false accusations can harm both accused people and genuine victims of harassment. This reduces the distance to 0.0206, which corresponds to a 65.2% improvement. With PBPO-lite, the prompt is rewritten as: *Unjust accusations against men regarding sexual misconduct*. The rewrite itself is short and still related to the original topic. However, after this rewritten prompt is sent to Llama, the generated answer shifts into a specific first-person narrative about being accused of sexual assault. This increases the distance to 0.0992, corresponding to a -67.4% change relative to Base.

This example helps explain why PBPO-lite can be less stable than URIAL. PBPO-lite may preserve the topic at the rewriting stage, but the rewritten prompt can still lead the downstream model into a more sensitive or less aligned response. URIAL does not rewrite the original prompt. Instead, it conditions the target model toward a safer and more general answer, which in this case keeps the output closer to the accepted PRISM answer.

6 Discussion

This project starts from a practical alignment problem: safety behaviour is often applied as one generic standard, even though users differ in how they perceive harmful wording [7; 11; 1]. At the same time, retraining or fine-tuning a model for every user is not realistic [10; 11]. This thesis therefore tested whether user-specific safety signals can be introduced before decoding, while keeping the target model fixed [10]. The results show that this is possible: both URIAL and PBPO-Lite reduce toxicity-distance, but they differ in how much they preserve the original task.

The strongest takeaway is that simple prompt-context steering is a competitive approach to personalized safety. URIAL performs best overall, especially on Llama, and it does so without changing the user's original prompt. This matters because it reduces the risk of altering the task before the model answers. In this experiment, adding a shared safety instruction and category-specific examples was enough to move outputs closer to preferred PRISM answers in toxicity-score space [8; 6]. This supports the broader idea behind training-free alignment: some alignment behaviour can be controlled at inference time through the input context rather than through model updates [9; 10].

PBPO-Lite shows a different trade-off. Its advantage is that it uses a richer user profile: instead of representing a user with one primary toxicity category, it uses six sensitivity dimensions. This makes it a more expressive form of personalization [11]. However, the sensitivity profile is applied by rewriting the user's prompt before the target model sees it [3]. The results show that this can reduce toxicity-distance, but also introduces a task-preservation problem. In particular, the MMLU results suggest that rewriting is risky when the exact wording, options, or format of the prompt matter. PBPO-Lite therefore appears more suitable for open-ended prompts that contain harmful wording than for structured knowledge tasks.

The auxiliary evaluations show why toxicity reduction should not be studied in isolation [4]. Perplexity decreases under both interventions, suggesting that the methods do not reduce toxicity-distance by making outputs less fluent under this automatic proxy. However, knowledge-task performance is more fragile [5]. Qwen with URIAL gives the best balance between toxicity-distance reduction and MMLU retention, while Llama is more sensitive to both interventions.

This means that pre-decoding alignment is model-dependent: a method that works well for one target model or task type may not transfer cleanly to another [10].

These findings also clarify what personalization means in this setup [7; 11]. The ablations suggest that part of the improvement comes from general safety steering, not only from user-specific profiles. However, user-specific components still help in some settings, especially when compared with generic or mismatched alternatives. This means the results should be interpreted as evidence for lightweight personalized steering, but not as proof that the current profile assignment is optimal. The way users are mapped to toxicity categories in URIAL, and to six-dimensional sensitivity labels in PBPO-Lite, matters for the final behaviour.

Overall, the results suggest that personalized pre-decoding is most promising when it gives the model useful safety context without unnecessarily changing the task [10; 11]. URIAL currently offers the cleaner trade-off because it preserves the original prompt. PBPO-Lite shows the potential of richer profile-based intervention, but also highlights the need for safeguards that check whether the rewrite still expresses the same intent [3]. Future personalized safety systems should therefore optimize not only for lower toxicity, but also for intent and task preservation, and robustness across models.

6.1 Limitations

This project has several limitations. First, personalized URIAL is simple to implement, but it is not very scalable. It depends on manually written examples for each toxicity-sensitivity category. Extending this to more categories, domains, or user-specific contexts would require more example design. The current version also assigns each user one primary sensitivity category, which keeps the method interpretable but loses information when a user is sensitive to multiple toxicity types at once.

Second, PBPO-Lite depends heavily on the rewriting model. Even when a rewrite makes a prompt less harmful, it can change the user's intent, remove information, or alter the task structure. This is especially problematic for factual and multiple-choice tasks, where small wording changes can affect the answer. The method would need stronger constraints or verification before it could be used reliably in settings where exact task preservation matters.

Third, the evaluation is limited by the dataset and metric. PRISM is useful because it links prompts, rated answers, user identifiers, and user information, making it possible to study personalization [8]. However, there are few comparable datasets, so it is hard to know whether the same conclusions would hold for other users or more explicitly toxic prompts. The toxicity-distance metric is also only an automatic proxy [6; 4; 2]. It measures movement toward accepted PRISM answers in Perspective-score space, not definitive human-perceived safety [1].

Finally, the experiments use two target models and one rewriting model. The model-dependent results show that the findings should not be treated as universal across all LLMs [10]. Different model families, model sizes, or instruction-tuning procedures may respond differently to the same pre-decoding intervention.

6.2 Future Work

Future work should evaluate personalized pre-decoding with stronger human evidence [7; 11]. A natural next step would be to run a dedicated user study and build a dataset focused specifically on personalized toxicity alignment, with more toxic or sensitive prompts than PRISM contains [8]. Human review would be important here because the current toxicity-distance metric only shows whether outputs move closer to preferred PRISM answers under an automatic scoring metric [6]. It does not show whether users actually perceive the responses as safer, more respectful, or better aligned with their preferences [1; 4; 2].

PBPO-Lite should be improved with explicit intent-preservation checks, especially for structured tasks such as multiple-choice questions. More broadly, future systems could learn when to apply prompt conditioning, when to rewrite the prompt, and when not to intervene at all. This would make personalized pre-decoding more adaptive while reducing the risk of unnecessary task changes.

The experiments should also be scaled to more prompts, more prompt-sampling seeds, more target models, and larger knowledge benchmarks. Testing beyond Qwen and Llama would make it clearer how broadly the results generalize across model families and whether the same trade-offs appear for larger or differently aligned models [10]. This would also help separate method-level effects from model-specific behaviour.

7 Conclusion

This project asked whether personalized pre-decoding alignment can reduce toxicity in LLM outputs without fine-tuning, while preserving fluency and usefulness. The answer is partly yes. Both tested methods reduce toxicity-distance compared with the base condition. URIAL is the strongest method, reaching about 31% mean improvement on Qwen and 51% on Llama across four PRISM prompt-sampling seeds. PBPO-Lite also improves toxicity-distance, but the gains are smaller and more variable.

The main contribution of this work is therefore not only that toxicity-distance can be reduced, but that this reduction comes with important trade-offs. Perplexity decreases for both methods on both target models, suggesting that the interventions do not make outputs less fluent under this proxy metric. However, MMLU results show that pre-decoding safety interventions can harm knowledge-task performance, especially when the prompt format is changed. This shows that personalized pre-decoding is practical because it does not require training the target model, but it must be designed carefully so that safety improvements do not interfere too much with the original task.

8 Responsible Research

This project studies toxicity reduction, so ethical reflection is central to the work. A personalized safety system can protect users better than a single generic standard, but it can also introduce risks. If a system over-personalizes, it may reinforce a user's existing biases or suppress legitimate expression [11]. For example, if a user consistently rates responses with strong

disagreement or direct political criticism as unsafe, a personalized system may learn to soften or remove such criticism even when it is legitimate. Similarly, if a user is unusually tolerant of insults toward a certain group, the system may under-correct harmful language for that user. In both cases, personalization could amplify the user's existing preferences instead of providing a responsible safety improvement. If the system neutralizes language too aggressively, it may also remove important political, cultural, or emotional meaning. For instance, a statement that reports discrimination, quotes harmful language for educational reasons, or uses reclaimed language within a community may be wrongly rewritten as if it were itself harmful. For that reason, the goal of this project is not to censor difficult topics. The goal is to reduce harmful wording while preserving the original user intent.

The use of PRISM also requires care. PRISM contains human feedback and user-linked information, so the experiments should avoid exposing personally identifiable information [8].In this work, user IDs are not treated as real names or identities, and the sensitivity profiles are used only for this experiment. The methods use sensitivity profiles derived from existing data rather than attempting to infer private beliefs beyond the dataset. This distinction is important because the sensitivity profile should only be interpreted as a task-specific signal about how a user rated previous model responses, not as a general psychological or political profile of that user.

The toxicity metric is another ethical limitation. Perspective-style scores are classifier outputs, not human judgments. They can miss context and may contain biases [1; 4; 2]. For example, a classifier may assign a high toxicity score to a response that discusses hate speech, quotes an offensive phrase in order to criticize it, or uses profanity to express distress rather than to attack another person. It may also behave unevenly across identity terms, dialects, or cultural language varieties, which can lead to higher toxicity scores for some communities or topics even when the intended meaning is not harmful. Therefore, the results should not be interpreted as final proof that an answer is safe. They show that the generated answer moved closer to the preferred PRISM answer in a specific toxicity-score space. A stronger evaluation would include human review, especially for cases where the text contains identity-related, political, cultural, or quoted harmful language.

Reproducibility is important because LLM experiments can change when prompts, model versions, random seeds, decoding settings, or scoring tools change. To make the work easier to reproduce, I used fixed prompt-sampling seeds, fixed target models, and the same toxicity-distance metric for all runs. The prompt templates are also fixed, including the URIAL safety prefixes, URIAL example selections, and PBPO-Lite rewrite instructions. I also stored the generated answers, toxicity scores, preferred-answer distances, and summary files, so the final results can be checked without repeating every generation step. There are still parts that cannot be reproduced exactly. For example, the generated answers may change if settings such as temperature, maximum output length, or sampling parameters are changed. External scoring tools may also change over time, so cached toxicity scores are important for checking the reported results.

8.1 Use of Generative AI

I used ChatGPT, Gemini, and Codex as supporting tools during this project for brainstorming, improving writing clarity and structure, code assistance, debugging, and documentation support.

All AI-assisted suggestions were reviewed, verified, and edited by me before being included. The tools were not used to replace my own research, reasoning, analysis, conclusions, or authorship. I remain fully responsible for the final content, code, results, interpretations, and academic integrity of this work.

Where relevant, external sources and ideas were cited appropriately, and no sensitive, confidential, proprietary, or personal data was entered into these tools without proper permission.

References

[1] Sergei Berezin, Reza Farahbakhsh, and Noel Crespi. Toxicity detection should measure contextual harm, not text-intrinsic badness. 2026. URL: https://arxiv.org/abs/2503.16072, doi:10.48550/arXiv.2503.16072.

[2] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*. Association for Computing Machinery, 2019. doi:10.1145/3308560.3317593.

[3] Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 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)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL: https://aclanthology.org/2024.acl-long.176/, doi:10.18653/v1/2024.acl-long.176.

[4] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, 2018. doi:10.1145/3278721.3278729.

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. 2020. URL: https://arxiv.org/abs/2009.03300, doi:10.48550/arXiv.2009.03300.

[6] Jigsaw. Perspective api. https://perspectiveapi.com/, 2026.

[7] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. 2023. URL: https://arxiv.org/abs/2303.05453, doi:10.48550/arXiv.2303.05453.

[8] 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 A. Hale. 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*, 2024. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html, doi:10.52202/079017-3342.

[9] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Raghavi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. 2023. URL: https://arxiv.org/abs/2312.01552, doi:10.48550/arXiv.2312.01552.

[10] Birong Pan, Yongqi Li, Weiyu Zhang, Wenpeng Lu, Mayi Xu, Shen Zhou, Yuanyuan Zhu, Ming Zhong, and Tieyun Qian. A survey on training-free alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, 2025. Association for Computational Linguistics. URL: https://aclanthology.org/2025.findings-emnlp.238/, doi:10.18653/v1/2025.findings-emnlp.238.

[11] Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, Prithviraj Ammanabrolu, and Julian McAuley. A survey on personalized and pluralistic preference alignment in large language models. 2025. URL: https://arxiv.org/abs/2504.07070, doi:10.48550/arXiv.2504.07070.