Personalized Pre-Decoding Alignment for Training-Free Toxicity Reduction

Comparing URIAL and PBPO-Lite on PRISM user prompts without fine-tuning

TU Delft

1 Introduction

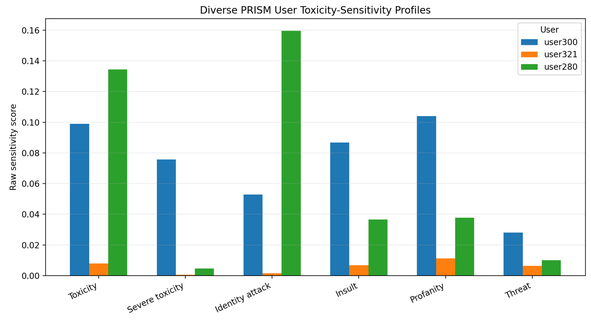
Toxicity is not one-size-fits-all. It is subjective and context-dependent — what offends one user may be acceptable to another. A single safety standard is too strict for some and too lax for others.

This motivates **personalized alignment within safe bounds**, yet fine-tuning a separate model per user is expensive and often impossible without weight access.

Training-free alignment adapts behaviour at inference time without changing weights, and can act at three stages:

pre-decoding → in-decoding → post-decoding

We focus on **pre-decoding**: modifying the prompt *before* generation — simple, cheap, and model-agnostic.



Three real PRISM users differ sharply in toxicity sensitivity across categories ⇒ one universal threshold cannot fit everyone.

2 Research Questions

Main: How effectively can personalized pre-decoding interventions align LLM outputs with user-specific toxicity preferences *without fine-tuning?*

SQ1 · Toxicity

How much do they reduce **toxicity-distance** versus an unmodified base model?

SQ2 · Knowledge

How do they affect **knowledge-task performance** (one-shot MMLU accuracy)?

SQ3 · Fluency

How do they affect **fluency**, measured by perplexity of the answer?

⋆ Contributions

1. A direct comparison of two training-free *personalization styles* — adding context vs. rewriting the prompt — for toxicity reduction.

2. A user-grounded **toxicity-distance** evaluation on PRISM, scored against each user's own preferred answer.

3. Evidence that safety gains must be balanced against **knowledge-task** cost, shown across two model families.

3 Methodology

Evaluated on **PRISM** user prompts, which link user profiles, prompts and preference feedback. Target models: **Qwen 2.5 7B** and **Llama 3.1 8B**. Two training-free interventions are compared against a **Base** condition.

Base: original prompt → target model → answer

URIAL · prompt conditioning

Keep the original prompt unchanged; prepend a shared safety instruction and **3 examples** chosen by the user's *primary* toxicity category.

instruction → personalized examples → original prompt

PBPO-Lite · prompt rewriting

A secondary model (**Mistral-7B**) rewrites the prompt from the user's **6-dimensional** toxicity profile, preserving the original topic and intent.

prompt + profile → Mistral rewrite → target model

4 Evaluation Metric

Toxicity-distance = mean absolute difference to the *preferred* PRISM answer across six Perspective dimensions (toxicity, severe toxicity, identity attack, insult, profanity, threat):

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

Reported as **percentage improvement** over Base (higher is better):

$$\mathrm{PI} = \frac{d_{\text{base}} - d_{\text{method}}}{d_{\text{base}}} \times 100$$

Knowledge retention is measured with **MMLU** (1000 one-shot questions); fluency with **perplexity**.

! The Rewriting Risk

Conditioning preserves the task. Rewriting is more expressive, but can **change the user's intent**:

Original: *"Gay people shouldn't get married."*

Rewrite: *"Marriage should be available to everyone, regardless of sexual orientation."*

The rewrite removes harmful framing but **reverses the stance** instead of neutralising it — a real failure mode of rewriting, and why URIAL is generally safer for task preservation.

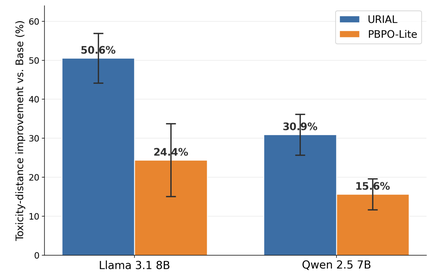
○ Experimental Setup

Dataset	PRISM — profiles, prompts & preferred answers
Sampling	4 seeds × 400 prompt–user pairs, 6 balanced categories
Models	Qwen 2.5 7B / Llama 3.1 8B; Mistral-7B rewriter
Scoring	Perspective API (6 dims) · MMLU 1000 Q · perplexity

5 Results

SQ1 · Toxicity-distance reduction

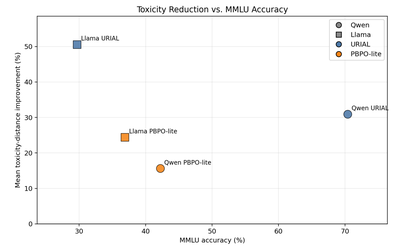
Both methods beat Base; **URIAL is strongest** on both models, and all four bootstrap **95% CIs lie fully above zero**.



Mean toxicity-distance improvement over four PRISM seeds; error bars show standard deviation across seeds.

SQ2 · Knowledge retention

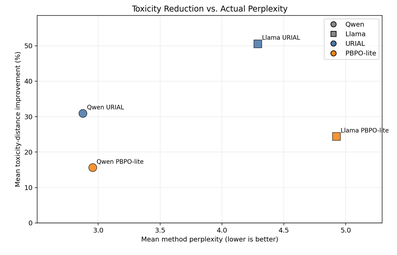
A clear **trade-off**: **Qwen + URIAL** keeps the best balance (70.4% vs. 72.4% Base); Llama + URIAL gives the largest toxicity drop but loses most MMLU.



Y: mean toxicity-distance improvement across PRISM seeds. X: MMLU accuracy from the 1000-question MMLU run.

SQ3 · Fluency (perplexity)

Both methods give **lower perplexity than Base** on both models ⇒ no fluency penalty under this proxy.



Y: mean toxicity-distance improvement across PRISM seeds. X: mean method answer perplexity across PRISM seeds; lower is better.

6 Conclusion

Personalized pre-decoding **reduces toxicity without fine-tuning**. **URIAL** gives the cleaner trade-off because it preserves the prompt; **PBPO-Lite** is more expressive but riskier, since rewriting can alter the task or intent.

Takeaway: safety gains must be weighed against knowledge-task losses — a promising, low-cost direction for personalized alignment.

Alina Florea | Supervisors: Anne Arzberger, Enrico Liscio | Responsible professor: Jie Yang | Examiner: Carolin Brandt

For references, see paper