

Identity-Preserving Detoxification via Local Causal Effect Ranking on Masked Diffusion Language Models

Yucheng Lin, Yue Liu, Jiaojiao Jiang

University of New South Wales

High St, Kensington NSW 2052, Australia

{yucheng.lin,yue.liu18,jiaojiao.jiang}@unsw.edu.au

Abstract

Text detoxification should produce safer rewrites while retaining identity mentions from the original text. However, classifier-guided methods can remove identity descriptors alongside genuinely harmful content when toxicity classifiers associate identity mentions with harmfulness and guidance relies on absolute scores. We propose CIDER (Contrastive Identity-preserving Detoxification via Explicit Reference), a default-relative, margin-gated selective-guidance method for masked diffusion language models. At each selected masked position, CIDER evaluates every candidate in the admissible top- k set relative to the model-default token in the same partially decoded state and commits a non-default candidate only when the best candidate lowers the one-step proxy score by more than a preset margin. On the primary LLaDA-8B CivilComments×Jigsaw setting, CIDER improves retention by +0.06 to +0.16 over the interpolated absolute-score frontier at matched residual-toxicity levels. At the reported aggressive endpoints, the retention advantage recurs across four data pairings on LLaDA-8B and three bases on CivilComments×Jigsaw. Controlled ablations indicate that the selection rule is an important contributor to the improvement.

Content warning. This paper describes toxic and discriminatory content that may be disturbing to some readers.

1 Introduction

Text detoxification is increasingly deployed in content moderation and conversational systems, requiring models to rewrite harmful content while preserving the protected identity-related mentions referenced in the original text. This requirement matters because identity groups appear in both toxic and non-toxic comments, yet modern toxicity classifiers tend to associate identity mentions

with toxicity in a statistical sense (Dixon et al., 2018; Borkan et al., 2019; Sap et al., 2022). The consequence is that a system treating identity mentions themselves as a toxicity signal can reduce its measured toxicity partly by removing the identity references under discussion.

Existing detoxification methods intervene at several loci. Some steer generation away from tokens that raise the toxicity score at decoding time (Dathathri et al., 2020; Krause et al., 2021; Liu et al., 2021) while others edit internal representations or parameters (Wang et al., 2024; Uppaal et al., 2025). These methods are effective when the score tracks harmful expression. In the absolute-score decoding setting studied here, however, guidance can act on identity-bearing context without testing whether the selected alternative improves the score relative to the model-default token in the same state. A complementary line debiases the toxicity classifier to reduce identity-linked score inflation (Lu et al., 2024; Xiao et al., 2025a). This reduces subgroup disparities in the scorer, but residual bias can remain. Even with a debiased scorer, a guidance-rule problem remains if generation still applies an absolute rule.

This tension suggests that detoxification guidance should change its target: a token should be suppressed only when changing it lowers the one-step predicted-harm proxy relative to the model-default token in the same partially decoded state, rather than when a sentence carrying an identity context receives a high score. Recent inference-time reward and representation-editing methods such as ARGRE (Xiao et al., 2025b) make step-wise intervention feasible, but an absolute reward signal does not condition the intervention on this fixed-state candidate–default contrast.

Motivated by this gap, we propose CIDER (Figure 1), a selective-guidance method for masked diffusion language models that replaces absolute-score suppression with a selective rule based on

a paired token-level score contrast. Rather than suppressing whenever a partially decoded sentence scores high, CIDER compares each candidate with the model-default token in the same partially decoded state, yielding a one-step contrast τ . The decoder replaces the default token only when the best candidate lowers the predicted-harm proxy beyond a threshold. We evaluate the resulting identity retention empirically across datasets and base models.

To make this practical, CIDER comprises three components: a one-step rewriter that embeds paired candidate scoring within decoding; a two-pass variant, CIDER Edit, which strengthens detoxification without reverting to absolute-score suppression; and distillation of the guided behaviour into a LoRA student so that inference cost returns to the base-model level.

On the primary LLaDA-8B setting with CivilComments paired with Jigsaw (Nie et al., 2025), CIDER’s retention gap over the interpolated absolute-score frontier is +0.06 to +0.16 at matched residual-toxicity levels. Complementary aggressive-endpoint comparisons reproduce the retention advantage across four data pairings on LLaDA-8B and across three masked-diffusion bases on CivilComments×Jigsaw (Ye et al., 2025; Zhu et al., 2025), with raw gaps of +0.54 to +0.80. On the primary setting, CIDER Edit retains 0.65 of identity mentions at residual toxicity 0.154, versus 0.11 for ABS@1.0 at residual toxicity 0.056. Across controlled ablations (budget-matched iteration, a reward×rule cross, and a soft-guidance control without paired differencing), the results indicate that the paired selection rule is an important contributor beyond extra compute, generic soft steering, or reward debiasing.

Our contributions are as follows:

- **Inspectable paired selective guidance.** We introduce a fixed-state candidate–default score contrast τ for masked-diffusion detoxification and use it to gate token replacement.
- **Identity preservation across settings.** We establish a retention advantage at matched detoxification on the primary setting, reproduce the aggressive-endpoint pattern across datasets and base models, and report a three-tier identity-preservation metric.
- **Mechanism analysis and distillation.** Controlled reward-by-rule and budget-matched

comparisons assess the role of the selection rule, and a LoRA student reduces inference overhead to base-model cost.

2 Related Work

Classifier-guided and preference-based detoxification. Most guided detoxification methods reduce toxicity at decoding time by steering the generation away from tokens that raise an external classifier’s score; representative work includes PPLM’s plug-and-play control (Dathathri et al., 2020), GeDi’s discriminator-guided decoding (Krause et al., 2021), and DExperts’ mixture of experts and anti-experts (Liu et al., 2021). A second line modifies the model itself, either through preference learning such as DPO (Rafailov et al., 2024) and VRPO (Zhu et al., 2025), or through parameter editing such as DINM (Wang et al., 2024) and ProFS (Uppaal et al., 2025). These approaches intervene at different stages and optimise different objectives. CIDER instead keeps the model frozen and modifies only the sampler, replacing absolute-score selection with a paired score contrast between a candidate and the model-default token in the same decoding state.

Inference-time representation steering. The closest neighbour is ARGRE (Xiao et al., 2025b), which steers internal activations toward non-toxic directions at test time; sparse-autoencoder guidance is a recent variant (Goyal et al., 2025). CIDER differs from these along three axes: it operates on the discrete masked-diffusion transition rule rather than on continuous hidden states; its signal comes from a fixed-state paired score contrast rather than an absolute reward score; and its objective is identity preservation rather than detoxification alone.

Fairness in toxicity classifiers. Subgroup disparities in toxicity scoring are well documented (Dixon et al., 2018; Borkan et al., 2019). Recent work mitigates them at several loci: at the evaluation side, by selectively invoking a fairness intervention based on a toxicity assessment (Ren et al., 2026); at the classifier side, through paired counterfactual minimal pairs (Lu et al., 2024); and at the representation side, through activation debiasing (Xiao et al., 2025a). CIDER studies the complementary generation-side decision of how a toxicity score triggers an edit. Controlled comparisons indicate that the selection rule remains an important contributor when the scorer is debiased.

3 Method

As illustrated in Fig. 1, CIDER comprises a frozen masked diffusion language model and harm reward (§3.1), a teacher rewriter that applies a fixed-state token intervention, paired score contrast, and selective rule (§3.2), and a LoRA student delivering detoxification at base-model inference cost without per-step reward scoring (§3.3).

3.1 Preliminaries

Masked Diffusion Language Model. We operate on a frozen masked diffusion language model p_θ , instantiated as LLaDA-8B-Instruct (Nie et al., 2025). A noisy sequence $\mathbf{x}_t \in \mathcal{V}^L$ is iteratively denoised toward $\mathbf{x}_0$. Our guided decoder selects one unresolved position per decoding iteration and writes one candidate token there, using the conditional distribution $p_\theta(x_i | \mathbf{x}_{t,-i})$ at position i . Here, $\mathcal{V}$ denotes the model vocabulary, L denotes the sequence length, and $m \in \mathcal{V}$ denotes the dedicated mask token; for LLaDA-8B, $|\mathcal{V}| = 126,464$. We write $\mathbf{x}_{t,-i}$ for the current partially decoded state with position i masked and $\mathbf{x}_{t,i \leftarrow c}$ for that state with position i forced to value c . For each base model, $\emptyset \neq \mathcal{V}_{\text{adm}} \subseteq \mathcal{V}$ denotes the model-specific candidate vocabulary used for position confidence, top- k , and default-token selection. We write its renormalised restriction as

$$p_\theta^{\text{adm}}(v | \mathbf{x}_{t,-i}) = \frac{\mathbf{1}\{v \in \mathcal{V}_{\text{adm}}\} p_\theta(v | \mathbf{x}_{t,-i})}{\sum_{w \in \mathcal{V}_{\text{adm}}} p_\theta(w | \mathbf{x}_{t,-i})}.$$

The construction of $\mathcal{V}_{\text{adm}}$ is detailed in App. B.

Harm Reward. We assume access to a text-level harm reward R_ϕ trained to predict toxicity. In our instantiation, R_ϕ is a RoBERTa-base encoder (Liu et al., 2019) with a scalar head trained on Civil Comments (Borkan et al., 2019) using within-group counterfactual data augmentation (CDA; Maudslay et al., 2019), which makes R_ϕ approximately invariant under within-group identity swaps. Throughout the selection rule, R_ϕ denotes the post-sigmoid toxicity probability, so $R_\phi(\cdot) \in [0, 1]$. We do not assume R_ϕ is unbiased. For an initialised decoding state, let $\mathcal{G}$ denote the appended generation-region indices and define

$$S_\theta(\mathbf{x}; \mathcal{G}) = \text{Decode}_\theta^{\text{skip-special}}((x_j)_{j \in \mathcal{G}, x_j \neq m}).$$

The composite scorer map first applies S_θ and then tokenises the resulting Unico text for the scorer with maximum input length ℓ :

$$\Psi_\ell(\mathbf{x}; \mathcal{G}) = \text{Tok}_\ell(S_\theta(\mathbf{x}; \mathcal{G})).$$

We suppress $\mathcal{G}$ when it is clear from the initialised state. The length is a fixed protocol parameter: ℓ_R denotes the value used by R_ϕ , and ℓ_{guide} is defined analogously for R_{guide} . The analysis in §3.2 studies the residual bias that can remain.

Over-Filtering Problem. Let $\text{ABS}@ \lambda$ denote the standard classifier-guided detoxifier with strength $\lambda \in [0, 1]$. At each generation step, let $s = (\mathbf{x}_t, i)$ denote the current state and selected masked position, let $\pi(i, s)$ be its top- k candidate set, and define

$$g(a; s) = R_{\text{guide}}(\Psi_{\ell_{\text{guide}}}(\mathbf{x}_{t,i \leftarrow a})),$$

where R_{guide} is either a stock toxicity classifier or the CDA-trained R_ϕ . Then $\text{ABS}@ \lambda$ selects

$$c_{\text{ABS}@ \lambda}^* = \begin{cases} \arg \min_{a \in \pi(i, s)} g(a; s) & \text{w.p. } \lambda, \\ c_0(s) & \text{w.p. } 1 - \lambda, \end{cases} \quad (1)$$

where $c_0(s)$ denotes the model’s default admissible token at position i . On benign inputs that mention protected identities, $\text{ABS}@ \lambda$ over-filters. Toxicity classifiers are statistically correlated with the mere mention of identity, so the $\arg \min g$ step preferentially removes identity tokens together with truly harmful ones. Our goal is a guidance procedure that achieves comparable detoxification while preserving identity language and fluency.

3.2 CIDER Teacher Rewriter

CIDER replaces the absolute-score selection rule in Eq. (1) with a default-relative score contrast and a selective margin gate. At a fixed decoding state, forcing a candidate token is a controlled intervention on the computational decoder; the outcome considered here is the resulting one-step scorer output.

Fixed-State Token Intervention. At a chosen position i during reverse diffusion, let $s = (\mathbf{x}_t, i)$ denote the pre-intervention partially decoded state. For each admissible candidate token a , we force $x_i := a$ and define the one-step scorer outcome

$$Y(a; s) = R_\phi(\Psi_{\ell_R}(\mathbf{x}_{t,i \leftarrow a})), \quad (2)$$

where both candidate evaluations use the same serialization and scorer. A full completion for every candidate would target a different, downstream outcome and would be substantially more expensive.

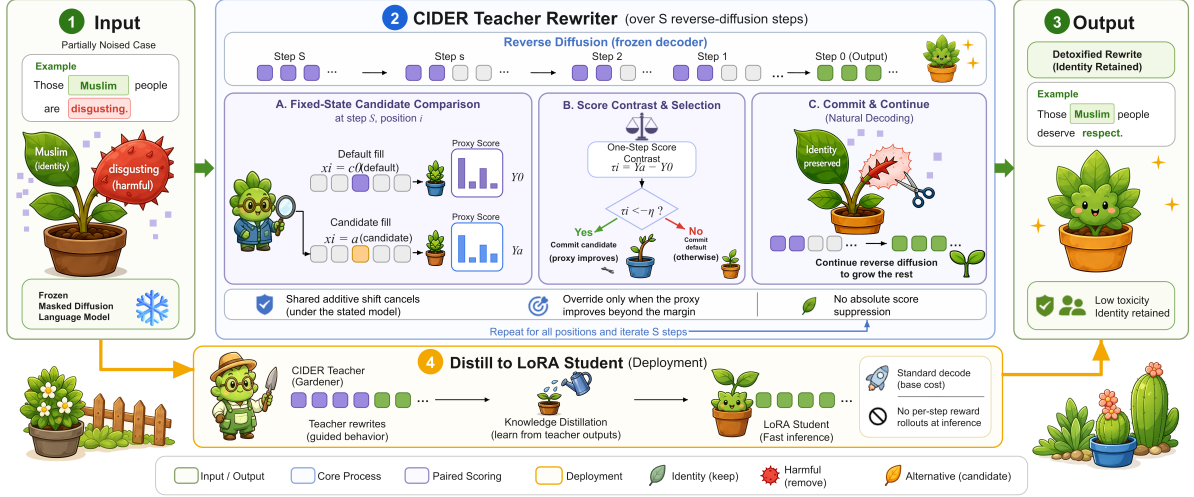


Figure 1: **Overview of CIDER.** (1) **Input:** a partially noised sequence in which a benign identity token (*Muslim*) co-occurs with a harmful token. (2) **Teacher rewriter:** at each reverse step, each candidate fill is compared with the model-default fill in the same partially decoded state; a non-default commit is made only when the best candidate $a_i^*(s)$ satisfies $\tau_i(a_i^*(s); s) < -\eta$. (3) **Output:** a detoxified rewrite that retains the identity reference. (4) **Deployment:** the behaviour is distilled into a LoRA student that decodes at base-model inference cost with no per-step reward scoring.

Default-Relative Score Contrast. Assume $1 \leq k \leq |\mathcal{V}_{\text{adm}}|$. We treat TopK_k as the ordered list returned by the decoder and use first-occurrence tie-breaking. Let

$$c_0(s) = \left[\text{TopK}_k \left(p_{\theta}^{\text{adm}}(\cdot \mid \mathbf{x}_{t,-i}) \right) \right]_1 \quad (3)$$

denote the base decoder’s default admissible token. For candidate a , define

$$\tau_i(a; s) = Y(a; s) - Y(c_0(s); s). \quad (4)$$

Because the pre-intervention state, serialization, and scorer are fixed, $\tau_i(a; s)$ is an exact state-specific interventional contrast on the computed one-step scorer outcome, relative to the base decoder’s default action.

Selective Rule and Iteration. Define

$$\pi(i, s) = \text{TopK}_k \left(p_{\theta}^{\text{adm}}(\cdot \mid \mathbf{x}_{t,-i}) \right), \quad (5)$$

so that $c_0(s) \in \pi(i, s)$, and let

$$a_i^*(s) = \arg \min_{a \in \pi(i, s)} Y(a; s). \quad (6)$$

The selective rule with threshold $\eta > 0$ commits

$$c_i^* = \begin{cases} a_i^*(s) & \text{if } \tau_i(a_i^*(s); s) < -\eta, \\ c_0(s) & \text{otherwise.} \end{cases} \quad (7)$$

Thus, the decoder departs from its default only when the best candidate in $\pi(i, s)$ lowers the one-step proxy score by more than η . Subtracting the default score calibrates this state-local gate; it does not change the candidate ranking because $Y(c_0(s); s)$ is constant across candidates at fixed s . Identity retention is evaluated as an empirical outcome of the rule.

Iterating Eqs. (2)–(7) over all reverse-diffusion steps yields a single guided generation. CIDER admits a *conservative* configuration (CIDER@0.20) with a narrow candidate set and a high threshold, which preserves base retention almost exactly while delivering moderate detoxification, and an *aggressive* configuration (agg) with a wider candidate set and a lower threshold that trades a small amount of retention for substantially larger detoxification. We further iterate either configuration P times, so the output of the first pass becomes the input to the second. This extends the observed detoxification range beyond the single-pass operating point without invoking an absolute-score rule. For the iterated variant, we use $P = 2$, and refer to the iterated aggressive configuration $\text{agg}_2 = \text{AGGRESSIVE} \circ \text{AGGRESSIVE}$ as *CIDER Edit*, our main two-pass procedure (Alg. 1). Full pseudocode and cross-family candidate handling are given in App. B.

Candidate-Independent Additive-Shift Analysis. Write $\bar{R}_\phi(\mathbf{x}) = R_\phi(\Psi_{\ell_R}(\mathbf{x}))$ for the composed scorer on a diffusion state. Suppose that, for every candidate a at fixed state s , the scorer decomposes as

$$\bar{R}_\phi(\mathbf{x}_{t,i \leftarrow a}) = b(\mathbf{x}_{t,-i}) + u_i(a; \mathbf{x}_{t,-i}). \quad (8)$$

Paired differencing gives

$$\tau_i(a; s) = u_i(a; \mathbf{x}_{t,-i}) - u_i(c_0(s); \mathbf{x}_{t,-i}), \quad (9)$$

so any candidate-independent additive term shared by the two scorer evaluations cancels. This is an algebraic shift-invariance property, not identification of a context-free harm contribution: candidate–context interactions and token-specific scorer bias may remain. The same shared shift also leaves within-state candidate rankings unchanged; CIDER differs operationally through its default-relative margin gate, not through a different ranking. We therefore evaluate the empirical stability of τ under context perturbations in App. A.

3.3 Distillation for Efficient Inference

Alg. 1 adds one batched R_ϕ forward pass per reverse-diffusion step on top of the base p_θ forward. With R_ϕ a RoBERTa-base (125M) scorer against the 8B-parameter LLaDA, this overhead is small: the conservative configuration costs $\approx 1.1 \times$ base inference and the two-pass CIDER Edit (agg₂, wider $k = 12$ over two passes) $\approx 2.4 \times$. This is acceptable for offline rewriting but roughly doubles serving latency.

To eliminate this overhead, we distil the guided behaviour into a LoRA student (Hu et al., 2021) $p_{\theta+\Delta_{\text{LoRA}}}$ on a held-out training pool $\mathcal{D}$. For $n = |\mathbf{y}^*|$, let $M = (M_1, \dots, M_n)$ be the target-position mask and define $\tilde{y}_j^{(M)} = m$ when $M_j = 1$ and $\tilde{y}_j^{(M)} = y_j^*$ otherwise. We draw $(\mathbf{u}, \mathbf{y}^*) \sim \mathcal{D}$ and $M_j \stackrel{\text{iid}}{\sim} \text{Bernoulli}(0.5)$, skipping draws with $\|M\|_1 = 0$. For each retained draw, the masked-diffusion loss is

$$\mathcal{L}_{\text{MD}}(\Delta; \mathbf{u}, \mathbf{y}^*, M) = -\frac{1}{\|M\|_1} \sum_{j=1}^n M_j \log p_{\theta+\Delta,j}(y_j^* | \mathbf{u}, \tilde{\mathbf{y}}^{(M)}).$$

The LoRA update minimises its expectation over these retained joint draws:

$$\Delta_{\text{LoRA}}^* = \arg \min_{\Delta \in \mathcal{A}_{\text{LoRA}}} \mathbb{E}[\mathcal{L}_{\text{MD}}(\Delta; \mathbf{u}, \mathbf{y}^*, M)]. \quad (10)$$

where $\mathcal{A}_{\text{LoRA}}$ is the chosen low-rank update family, $\mathbf{y}^* = \text{CIDERREWRITE}(\mathbf{u}, P)$ is the teacher output, and $p_{\theta+\Delta,j}$ is the student’s distribution at target position j . The expectation is over the retained draws specified above. Thus the student predicts masked target tokens from the prompt and unmasked target context, with cross-entropy averaged over the masked positions. This uses the same masked-token prediction form as the base p_θ objective. At inference, the student uses the ordinary masked-diffusion decoding schedule of the base model with no per-step R_ϕ rollouts, so the candidate-rollout overhead is removed and inference uses only the LoRA-adapted base model.

4 Experiments

We first describe the experimental setup and validate the identity-retention metric against independent reviewers (§4.1). We then report the matched frontier on the primary setting (§4.2), examine the selection rule through controlled comparisons (§4.2), and test whether the aggressive-endpoint pattern recurs across datasets and base models (§4.3). Finally, we compare against the broader baseline set (§4.4) and distil CIDER into a LoRA student (§4.5). Qualitative comparisons, context-perturbation analysis, and step-level traces are in App. H, App. A, and App. B.

4.1 Experimental Setup

Models. We evaluate CIDER on three frozen masked-diffusion language models: LLaDA-8B-Instruct (Nie et al., 2025) (primary), LLaDA-1.5 (Zhu et al., 2025) (within-family cross-version), and Dream-7B-Instruct (Ye et al., 2025) (cross-family). All bases run in bf16 on a single A100-40GB.

Datasets. We evaluate on the *Jigsaw Unintended Bias in Toxicity Classification* release of CivilComments (Borkan et al., 2019; cjadams et al., 2019), which provides 24 identity-subgroup annotations per comment. Around the goal of identity-preserving detoxification, we construct four complementary (benign, toxic) evaluation pairings: CivilComments benign inputs paired with Jigsaw-Identity, ToxicChat, or ToxiGen toxic inputs, and Dixon-Madlibs benign templates paired with Jigsaw-Identity. The CivilComments benign pool contains 1,639 held-out comments that mention a protected identity yet are labelled non-toxic, used to probe over-filtering; the toxic pools con-

tain Jigsaw-Identity ($N = 294$), ToxicChat (Lin et al., 2023) ($N = 746$), and ToxiGen (Hartvigsen et al., 2022) ($N = 402$), spanning platform comments to real human-machine dialogue. Dixon-Madlibs (Dixon et al., 2018) provides the synthetic cross-distribution benign pool, while HolisticBias (Smith et al., 2022) contributes identity descriptors to the evaluation vocabulary (App. C). All evaluation corpora share a single source-independent hash split over the normalised comment text (buckets 0–7 train, 8 dev, 9 test); the guidance reward model is trained only on train buckets, so the same comment falls in the same split across all corpora and no evaluation sample leaks into R_ϕ . After a character-length filter of $[10, 512]$, for each seed $s \in \{1, \dots, 5\}$ we independently subsample $n_{\text{benign}} = 250$ and $n_{\text{toxic}} = 150$ per side, using the same pair across all configurations within a seed so that per-seed comparisons are paired.

Metrics. Detoxification quality is inherently a multi-objective trade-off, so we characterise a method along two complementary axes: identity retention and residual toxicity. Identity retention measures whether the protected identity references in the input survive in the rewrite. Because an identity can be preserved either verbatim or through an in-axis paraphrase (e.g. “gay” $\rightarrow$ “homosexual”, “Muslim” $\rightarrow$ “Islamic”), purely lexical matching underestimates true retention; we therefore design a three-tier graded metric—from whether any identity term appears (LEXICAL), to whether the same identity axis is preserved (SAME-AXIS), to whether all identity axes of the input are covered (ALL-AXIS). We take SAME-AXIS-KEPT as the primary metric, as it most closely matches the semantic intuition that “the group is still referenced”. We exclude 14 ambiguous tokens whose dominant sense is non-identity in CivilComments-style text from the retention metric (criteria and list in App. C). Residual toxicity is judged by a held-out ensemble of classifiers to reduce dependence on the idiosyncrasies of any single scorer. The reward model used for guidance and the classifiers used for evaluation are mutually independent, reducing direct coupling between guidance and evaluation. Unless stated otherwise, we report 5-seed mean $\pm$ standard deviation.

Reward model. The guidance signal R_ϕ is the CDA-trained RoBERTa-base scorer of §3.1; the guidance reward and the evaluation classifiers are mutually independent. On the held-out split R_ϕ

attains AUC 0.95, and CDA training lowers the matched-recall identity-bias disparity from 0.110 (stock classifier) to 0.035 without sacrificing discriminative power. CIDER’s reward-by-rule comparison tests whether the paired rule’s retention ordering persists across the stock and CDA-trained scorers (§4.2).

Baselines. We compare against three families. *Same-base, training-free:* the uninvolved LLaDA base, a WordFilter (LDNOOBW¹) decoder, and a fairness system prompt. *Same-base, training-time:* LoRA SFT and DPO on ParaDetox (Logacheva et al., 2022), isolating sampler- versus model-modification. *Cross-paradigm reference:* ProFS (Uppaal et al., 2025) on GPT-2 and a prompted Mistral-7B (Jiang et al., 2023), included for context only. The principal comparison is the classifier-guided absolute-score baseline ABS@ λ , which at each step selects the score-minimising token with probability λ ; it shares CIDER’s guidance framework and differs only in the selection rule, providing the controlled comparison central to this paper.

Implementation. All masked-diffusion systems use the same decoding protocol and paired test inputs. We fix CIDER hyperparameters on a disjoint development split before multi-seed test evaluation, with no per-cell tuning. Model-specific candidate handling and decoding pseudocode are reported in App. B; threshold sensitivity is reported in App. F.

Metric validity. We validate the primary SAME-AXIS-KEPT metric against two independent reviewers on a 300-pair stratified anchor (App. D). Cohen’s κ between the metric and each reviewer is 0.726/0.738 (substantial); inter-reviewer κ is 0.874 on the 285 commonly-labelled pairs (almost-perfect). Both reviewers agree CIDER retains 67–90% of identity descriptors versus 10–14% for ABS@1.0.

4.2 Main Results and Ablations

Matched frontier on the primary setting. We compare CIDER against the classifier-guided baseline on identical inputs. On the primary LLaDA-8B CivilComments $\times$ Jigsaw setting, the three reported CIDER operating points lie above the base-anchored, piecewise-linear ABS@ λ frontier when

¹<https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words>

Table 1: Main frontier on LLaDA-8B, CivilComments×Jigsaw. All numbers are 5-seed mean $\pm$ std ($n_{\text{benign}} = 250$, $n_{\text{toxic}} = 150$ per seed). Retention is SAME-AXIS-KEPT; toxic_resid is the held-out three-classifier ensemble. At CIDER’s three reported residual-toxicity levels, the retention advantage over the interpolated ABS@ λ frontier is +0.06, +0.12, and +0.16, respectively (App. F).

config	retention $\uparrow$	toxic_resid $\downarrow$	detox $\uparrow$
base	0.791 $\pm$ 0.021	0.336 $\pm$ 0.014	+0.293
ABS@0.33	0.718 $\pm$ 0.007	0.236 $\pm$ 0.017	+0.393
ABS@0.66	0.346 $\pm$ 0.043	0.101 $\pm$ 0.006	+0.529
ABS@1.0	0.112 $\pm$ 0.013	0.056 $\pm$ 0.001	+0.573
CIDER@0.20	0.787 $\pm$ 0.021	0.252 $\pm$ 0.013	+0.378
agg	0.712 $\pm$ 0.023	0.192 $\pm$ 0.017	+0.438
agg ₂	0.649 $\pm$ 0.026	0.154 $\pm$ 0.012	+0.476
WordFilter	0.743 $\pm$ 0.038	0.324 $\pm$ 0.015	+0.306
Prompt	0.847 $\pm$ 0.031	0.495 $\pm$ 0.014	+0.135

retention is compared at the same residual-toxicity level.

Table 1 reports the primary frontier. At matched residual toxicity, computed by linearly interpolating the five-seed mean operating points on the base-anchored ABS frontier, CIDER’s retention advantage is +0.06 at $\rho = 0.252$, +0.12 at $\rho = 0.192$, and +0.16 at $\rho = 0.154$ (App. F). The conservative CIDER@0.20 configuration retains 0.787 identity descriptors at residual toxicity 0.252, while the two-pass aggressive agg₂ configuration (CIDER Edit) retains 0.649 at residual toxicity 0.154. For reference, the strongest absolute-score endpoint retains 0.112 at the lower residual toxicity of 0.056; the raw retention difference against this unmatched endpoint is +0.537 and is positive in all five seeds (App. F). WordFilter changes residual toxicity only slightly from the base (0.324 versus 0.336), while the fairness prompt raises retention but also raises residual toxicity to 0.495.

We examine the selection rule through three complementary controls: a budget-matched comparison tests whether an additional decoding pass is sufficient; a reward $\times$ rule cross tests whether the retention ordering persists across reward models; and a soft-steering sweep tests an absolute-score tilt without paired differencing. Together, these comparisons support the paired selection rule as an important contributor to identity preservation in the tested settings.

Budget-matched compute. On a subset of 40 benign-identity and 40 toxic inputs, we match the number of decoding passes between CIDER and

the absolute-score baseline. At one pass, CIDER retains 0.60 identity versus 0.05 for the absolute rule; at two passes the corresponding rates are 0.50 and 0.00. The absolute rule reaches somewhat stronger detoxification at both endpoints, but adding its second pass does not close the retention gap on this subset (App. G).

Reward $\times$ rule. To separate the retention patterns associated with the selection rule and reward debiasing, we run all four cells of {stock-classifier, CDA R_ϕ } $\times$ {absolute, paired} on the same inputs as Table 1. The paired rule retains 0.787 $\pm$ 0.021 identity with R_ϕ and 0.767 $\pm$ 0.019 with the stock classifier. The absolute rule retains 0.112 $\pm$ 0.013 with the stock classifier and 0.143 $\pm$ 0.028 with R_ϕ . The same ordering under both rewards supports a contribution from the selection rule, with reward debiasing providing a complementary change.

Soft steering without paired differencing. Finally, we test a generic inference-time detoxification bias without paired differencing. A soft absolute-score baseline STEER lowers each candidate’s logit by αR_ϕ , tilting persistently toward low-toxicity outputs, but uses no paired differencing. Sweeping α on LLaDA-1.5 (Table 13, App. G), STEER reaches residual $\rho \approx 0.256$ at $\alpha=24$ (−0.004 from $\alpha=16$), whereas CIDER agg₂ reaches $\rho = 0.165$, outside the range reached by STEER in the tested sweep. Thus, within $\alpha \in \{2, 4, 8, 12, 16, 24\}$, increasing the absolute-score tilt does not reach CIDER’s aggressive residual-toxicity endpoint.

4.3 Robustness

We test whether the aggressive-endpoint pattern recurs along two axes: across four (benign, toxic) distribution pairings on LLaDA-8B (CivilComments×{Jigsaw, ToxicChat, ToxiGen} and Dixon×Jigsaw) and across three masked-diffusion bases on CivilComments×Jigsaw (LLaDA-8B, LLaDA-1.5, and Dream-7B). At the reported endpoints, agg₂ retains substantially more identity language in every setting, with raw retention gaps of +0.54 to +0.80 (5-seed mean, std ≤ 0.03 throughout), while ABS@1.0 reaches lower residual toxicity. Table 15 in App. G reports the full breakdown.

Table 2: Full system comparison on CivilComments×Jigsaw. All systems are evaluated on the same pool (5-seed, $n_{\text{benign}} = 250$, $n_{\text{toxic}} = 150$ per seed); all numbers are 5-seed mean $\pm$ std. Retention is SAME-AXIS-KEPT; toxic_resid is the held-out three-classifier ensemble; PPL is the GPT-2-large median; infer is per-token cost relative to a base LLaDA-8B forward (1.0× denotes an exact single base forward; $\sim$ marks the approximate overhead of the paired-decode rows).

system	retention	toxic_resid	PPL	infer	paradigm
<i>Same-base, inference-time, no training</i>					
base	0.791 ± 0.021	0.336 ± 0.014	23.7	1.0×	LLaDA, no intervention
WordFilter	0.743 ± 0.038	0.324 ± 0.015	32.4	1.0×	LLaDA + dict
Prompt (fairness-aware)	0.847 ± 0.031	0.495 ± 0.014	29.5	1.0×	LLaDA + prompt
<i>Same-base, training-time (LLaDA-8B + LoRA)</i>					
SFT [†]	0.754 ± 0.057	0.374 ± 0.025	51.4	1.0×	LLaDA + train
DPO (ParaDetox prefs)	0.368 ± 0.064	0.147 ± 0.017	41.9	1.0×	LLaDA + train
<i>Cross-paradigm reference (autoregressive bases)</i>					
ProFS on GPT-2	0.509 ± 0.066	0.214 ± 0.024	5.2*	n/a	AR
Mistral-7B (prompted) [‡]	0.984 ± 0.016	0.136 ± 0.019	15.5	n/a	AR
<i>Ours (CIDER on LLaDA-8B)</i>					
CIDER@0.20	0.787 ± 0.021	0.252 ± 0.013	24.9	$\sim 1.1\times$	LLaDA + paired decode
CIDER agg₂	0.649 ± 0.026	0.154 ± 0.012	24.7	$\sim 2.4\times$	LLaDA + paired decode
CIDER distilled	0.754 ± 0.021	0.178 ± 0.010	31.7	1.0×	LLaDA + LoRA inference

[†] validation-selected SFT configuration from the reported LoRA grid; * GPT-2-large PPL on GPT-2-family outputs has known self-scoring bias; [‡] Mistral instruction-prompted.

4.4 Baseline Comparison

We compare CIDER with same-base baselines that share the LLaDA-8B backbone and with cross-paradigm references that use autoregressive bases.

Table 2 compares retention, residual toxicity, fluency, and per-token cost across all systems. WordFilter changes residual toxicity only slightly from the base and the fairness prompt raises it; their PPL is 24–37% above the LLaDA base. Under our ParaDetox construction, SFT frequently copies or truncates the input (App. E), yielding 0.754 retention but base-level residual toxicity. DPO reaches slightly lower residual toxicity than CIDER Edit (0.147 versus 0.154), whereas CIDER Edit retains substantially more identity language (0.649 versus 0.368) and has lower PPL (24.7 versus 41.9). CIDER@0.20 and CIDER Edit remain close to base fluency at PPL 24.9 and 24.7, respectively.

Prompted Mistral-7B attains the strongest numerical retention/residual pair in the table (0.984/0.136). Because it changes both the base model and decoding paradigm, we treat it as a cross-paradigm reference rather than a controlled test of the paired selection rule.

4.5 Distillation

The distilled LoRA student removes per-step reward rollouts and returns to 1.0× per-token inference cost. On LLaDA-8B it attains 0.754 retention and 0.178 residual toxicity, compared with

0.649/0.154 for the teacher and 0.791/0.336 for the base. The student therefore shifts toward higher retention and slightly weaker detoxification than the teacher. On LLaDA-1.5, student and teacher retention differ by 0.002 and residual toxicity by 0.009 (Table 12, App. G). We obtain the student by distilling $\approx 5,000$ teacher rewrites ($\mathbf{u}$, CIDERREWRITE($\mathbf{u}$, $P=2$)) from the CivilComments training split (identity-bearing; disjoint from evaluation by hash) following Eq. (10) of §3.3.

The LLaDA-8B student has PPL 31.7, compared with 24.7 for the teacher, and BERTScore against the input of 0.911, compared with 0.890 for the teacher. Distillation on Dream-7B requires base-specific adaptation left to future work. Context-perturbation analysis ($r = 0.87$), qualitative comparisons, and a failure-mode trace are in App. A, App. H, and App. B.

5 Conclusion

We showed that absolute-score detoxification on masked-diffusion language models can over-filter, removing identity tokens together with toxic ones when the guidance score is elevated by an identity-bearing context. CIDER instead applies a thresholded paired score contrast τ between a candidate and the model-default token in the same partially decoded state. Under an idealised additive decomposition, this differencing cancels the candidate-

independent additive context component shared by the two evaluations. Iterated into CIDER Edit and distilled into a LoRA student that runs at base inference cost, the method places all three reported operating points above the interpolated absolute-score frontier on the primary LLaDA-8B CivilComments×Jigsaw setting, with matched-retention advantages of +0.06 to +0.16. At the reported aggressive endpoints, higher retention recurs across four data pairings on LLaDA-8B and three bases on CivilComments×Jigsaw. Controlled budget-matched and reward-by-rule comparisons support the selection rule as an important contributor to identity preservation.

Limitations

Our approach has the following known boundaries.

(1) **Scope of token-level intervention.** CIDER operates at the token level and is therefore limited when the harm is carried by phrasal structure rather than an individual token (e.g. “white supremacist”), a case that can require semantic restructuring beyond the scope of this work.

(2) **Scope of the paired contrast.** Proposition 1 characterises an exact fixed-state difference between two one-step proxy evaluations. It is local and policy-relative; it is not a causal effect on the fully decoded rewrite or on human-perceived harm.

(3) **Operating regime.** CIDER targets the moderate-to-strong detoxification regime; under very mild detoxification on synthetic templates its advantage is not guaranteed, and our claims are restricted to that regime.

(4) **Paradigm scope.** Our method targets masked diffusion language models. Transferring the paired selection principle to autoregressive models involves different decoding semantics and is left to future work.

Ethical Considerations

This work studies identity-preserving detoxification and therefore processes toxic and discriminatory language. CIDER is a rewriting component for text already selected by a moderation or writing-assistance policy, not an autonomous arbiter of permissible speech. Experiments use only existing public datasets (Civil Comments (Borkan et al., 2019), Jigsaw (cjadams et al., 2019), ToxicChat (Lin et al., 2023), ToxiGen (Hartvigsen et al., 2022), and HolisticBias (Smith et al., 2022)); no

new toxic content was collected, and the paper includes a content warning.

The protected identity scope encoded by the evaluation vocabulary and metric is normative, and attack-frame cases remain a limitation (App. H). Two independent NLP reviewers were warned about toxic material, participated voluntarily, and could stop at any time. Like other detoxification systems, CIDER could be repurposed to amplify toxicity; inspectable per-token τ traces do not eliminate this risk.

Reproducibility. Decoding pseudocode and model-specific candidate handling are reported in App. B, dataset splits and identity vocabularies in App. C, and metric-validity anchors and the human-anchor rubric in App. D.

Acknowledgments

This work was supported by the Australian Research Council (ARC) under the DECRA Fellowship DE240101049. This research includes computations using the computational cluster Katana, supported by Research Technology Services at UNSW Sydney (PVC (Research Infrastructure), UNSW Sydney, 2010).

References

- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. [Nuanced metrics for measuring unintended bias with real data for text classification](#). *Preprint*, arXiv:1903.04561.
- cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification. <https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification>. Kaggle.
- 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](#). *Preprint*, arXiv:1912.02164.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. [Measuring and mitigating unintended bias in text classification](#). In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, page 67–73, New York, NY, USA. Association for Computing Machinery.
- Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. 2025. Breaking bad tokens: Detoxification of llms using sparse autoencoders. In *Proceedings of the 2025 Conference on*

- Empirical Methods in Natural Language Processing*, pages 12702–12720.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- 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. Association for Computational Linguistics.
- Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. [ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore. Association for Computational Linguistics.
- 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 (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. [ParaDetox: Detoxification with parallel data](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.
- Junyu Lu, Bo Xu, Xiaokun Zhang, Kaiyuan Liu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Take its essence, discard its dross! debiasing for toxic language detection via counterfactual causal effect. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15566–15578.
- Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. [Large language diffusion models](#). *Preprint*, arXiv:2502.09992.
- PVC (Research Infrastructure), UNSW Sydney. 2010. [Katana](#).
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- Jing Ren, Bowen Li, Ziqi Xu, Renqiang Luo, Shuo Yu, Xin Ye, Haytham Fayek, Xiaodong Li, and Feng Xia. 2026. When to invoke: Refining llm fairness with toxicity assessment. In *Proceedings of the ACM Web Conference 2026*, pages 9275–9285.
- 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. Association for Computational Linguistics.
- Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. ["i’m sorry to hear that": Finding new biases in language models with a holistic descriptor dataset](#). *Preprint*, arXiv:2205.09209.
- Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2025. Model editing as a robust and denoised variant of dpo: A case study on toxicity. In *International Conference on Learning Representations*, volume 2025, pages 69122–69153.
- Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi

Yang, Jindong Wang, and Huajun Chen. 2024. [Detoxifying large language models via knowledge editing](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3093–3118, Bangkok, Thailand. Association for Computational Linguistics.

Yisong Xiao, Aishan Liu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. 2025a. Fairness mediator: Neutralize stereotype associations to mitigate bias in large language models. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):250–273.

Yisong Xiao, Aishan Liu, Siyuan Liang, Zonghao Ying, Xianglong Liu, and Dacheng Tao. 2025b. [Detoxifying large language models via autoregressive reward guided representation editing](#). *Preprint*, arXiv:2510.01243.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. [Dream 7b: Diffusion large language models](#). *Preprint*, arXiv:2508.15487.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. [Llada 1.5: Variance-reduced preference optimization for large language diffusion models](#). *Preprint*, arXiv:2505.19223.

A Scope of the Fixed-State Intervention and Additive-Shift Analysis

Computational intervention. Fix a pre-intervention state $s = (\mathbf{x}_t, i)$. The treatment is the admissible token forced at position i , and the outcome is the one-step scorer value $Y(a; s) = R_\phi(\Psi_{\ell_R}(\mathbf{x}_{t,i \leftarrow a}))$. The reference action is the base decoder’s default admissible token $c_0(s)$.

Proposition 1 (Fixed-state one-step scorer intervention). *With the serialization and scorer fixed and the forced-token implementation given by Eq. (2),*

$$\tau_i(a; s) = Y(a; s) - Y(c_0(s); s)$$

is the exact, controlled state-specific contrast between forcing a and $c_0(s)$ on the one-step scorer outcome because both computational outcomes are queried at the same state. Its scope is the one-step scorer outcome; final-rewrite quality and identity retention are evaluated empirically.

Candidate-independent additive shift. Write $\bar{R}_\phi(\mathbf{x}) = R_\phi(\Psi_{\ell_R}(\mathbf{x}))$. If, for every candidate a ,

$$\bar{R}_\phi(\mathbf{x}_{t,i \leftarrow a}) = b(\mathbf{x}_{t,-i}) + u_i(a; \mathbf{x}_{t,-i}),$$

then paired differencing gives

$$\tau_i(a; s) = u_i(a; \mathbf{x}_{t,-i}) - u_i(c_0(s); \mathbf{x}_{t,-i}).$$

Proposition 2 (Candidate-independent additive-shift invariance). *Adding the same term $q(\mathbf{x}_{t,-i})$ to both candidate evaluations leaves $\tau_i(a; s)$ unchanged. Candidate–context interactions and token-specific scorer bias need not cancel. The proposition is an algebraic shift-invariance result; it does not identify separate harm and bias components of the scorer.*

Context-perturbation stability of the paired contrast. Proposition 2 removes only a term shared within each candidate–reference pair; it does not imply that the remaining contrast is invariant when the surrounding context changes. We therefore test empirical stability rather than exact additivity. For 80 identity-bearing inputs (per seed; 3 seeds), we (i) locate the identity-word position i , (ii) draw 4 candidate same-axis substitutions, and (iii) compare the original candidate–reference contrast with the corresponding contrast under 8 random permutations of $\mathbf{x}_{t,-i}$, preserving position i while reordering the surrounding words.

The median τ standard deviation under context permutation is 0.010 ($p_{90} = 0.038$). For comparison, the absolute contrasts at the firing decisions in the four reported, hand-picked traces are approximately 0.05–0.5 (App. B). The 3-seed mean Pearson r between the original τ and the permuted-mean τ is 0.87, and sign agreement is 84%. These results indicate stability under the tested perturbations. The residual 16% sign disagreement (with one seed at $r = 0.75$) also shows that exact permutation invariance is neither observed nor claimed.

B Algorithm and Decoding Traces

B.1 Pseudocode

For a partially decoded sequence $\mathbf{x}$ with generation region $\mathcal{G}$, let $\mathcal{M}_{\mathcal{G}}(\mathbf{x}) = \{j \in \mathcal{G} : x_j = m\}$ denote its unresolved generation positions, and define the base confidence at position j as

$$h(j; \mathbf{x}) = \max_{v \in \mathcal{V}_{\text{adm}}} p_\theta^{\text{adm}}(v \mid \mathbf{x}_{-j}).$$

Table 3: Context-perturbation stability of the paired contrast (3 seeds, 80 inputs each, 4 candidates per input, 8 random context permutations). Small τ standard deviation and high Pearson correlation indicate that the candidate contrast is empirically stable, but not invariant, under these perturbations.

seed	τ -std median	τ -std p90	Pearson r	sign agree
0	0.009	0.038	+0.949	0.82
1	0.012	0.039	+0.752	0.84
2	0.010	0.038	+0.899	0.87
mean	0.010	0.038	+0.867	0.84

Ordered top- k uses the order returned by the implementation; $\arg \max$ and $\arg \min$ use first-occurrence tie-breaking.

Algorithm 1 CIDER decoding; $P = 2$ with the aggressive settings gives CIDER Edit.

Require: Input $\mathbf{u}$; frozen p_θ and R_ϕ ; serialization Ψ_{ℓ_R} ; admissible vocabulary $\mathcal{V}_{\text{adm}}$; generation budget $L_{\text{gen}}; k, \eta, P$.

- 1: **function** CIDERREWRITE($\mathbf{u}, P$)
- 2: $\mathbf{y} \leftarrow \mathbf{u}$
- 3: **for** pass = 1, $\dots$, P **do**
- 4: $\mathbf{y} \leftarrow \text{GUIDEDDECODE}(\mathbf{y})$
- 5: **end for**
- 6: **return** $\mathbf{y}$
- 7: **end function**
- 8: **function** GUIDEDDECODE($\mathbf{u}$)
- 9: $(\mathbf{x}, \mathcal{G}) \leftarrow \text{INITIALISE}(\mathbf{u}, L_{\text{gen}}) \triangleright \text{prompt}$
+ masked generation region
- 10: **for** $r = 1, \dots, L_{\text{gen}}$ **do**
- 11: **if** $\mathcal{M}_{\mathcal{G}}(\mathbf{x}) = \emptyset$ **then**
- 12: **break**
- 13: **end if**
- 14: $i \leftarrow \arg \max_{j \in \mathcal{M}_{\mathcal{G}}(\mathbf{x})} h(j; \mathbf{x})$
- 15: $s \leftarrow (\mathbf{x}, i)$
- 16: $\pi(i, s) \leftarrow \text{TopK}_k(p_\theta^{\text{adm}}(\cdot | \mathbf{x}_{-i}))$
- 17: $c_0(s) \leftarrow [\pi(i, s)]_1$
- 18: Evaluate $Y(a; s) \leftarrow R_\phi(\Psi_{\ell_R}(\mathbf{x}_{i \leftarrow a}))$
- 19: **for all** $a \in \pi(i, s)$ **in one batch**
- 20: $\tau_i(a; s) \leftarrow Y(a; s) - Y(c_0(s); s)$ **for**
- 21: **all** $a \in \pi(i, s)$
- 22: $a_i^*(s) \leftarrow \arg \min_{a \in \pi(i, s)} Y(a; s)$
- 23: **if** $\tau_i(a_i^*(s); s) < -\eta$ **then**
- 24: $c^* \leftarrow a_i^*(s)$
- 25: **else**
- 26: $c^* \leftarrow c_0(s)$
- 27: **end if**
- 28: $\mathbf{x} \leftarrow \mathbf{x}_{i \leftarrow c^*}$
- 29: **end for**
- 30: **return** $S_\theta(\mathbf{x}; \mathcal{G})$
- 31: **end function**

Cross-family adaptation. Dream-7B uses a left-shifted logit convention: the raw output at position p predicts the token at position $p + 1$. We therefore realign logits by one position before selecting i , $\pi(i, s)$, and $c_0(s)$, while preserving the first-position boundary. For Dream-7B, $\mathcal{V}_{\text{adm}}$ excludes $\langle \text{mask} \rangle$, $\langle \text{endof text} \rangle$, and $\langle \text{pad} \rangle$ before both top- k and default-token selection. LLaDA-8B and LLaDA-1.5 instead use direct position-wise indexing.

B.2 Decoding Traces

We record per-step (position, top- k , R_ϕ , τ , decision) traces from CIDER-agg (one pass, $k = 12$, $\eta = 0.05$) on four hand-picked inputs and summarise them below.

$\diamond$ The race-term loss in Demo 2 is *not* a CIDER intervention — the rule fires 0 times. The base LLaDA distribution simply does not propose “black” or “white” in its top- k at the relevant positions; CIDER cannot inject tokens the base never considers. This is a recall failure of the base, not an over-filtering failure of guidance, and is consistent with the fixed-state scope of Prop. 1: the comparison and gate operate only over $\pi(i, s)$.

Detailed step (Demo 3, “white supremacist”): an identity-entangled failure case. At step 23 (position 57, near the “alt-right” token), the top-12 candidates and their τ values are

```
{-right: 0, alt: +9e-4, white: +0.046,
nationalist: +0.013, ist: +0.002,
ac: +0.001, extremist: +0.075, racist:
+0.481, supremacy: +0.012, far: +0.002,
suprem: -1e-4, right: -5e-4}.
```

The listed candidates white, racist, supremacy, nationalist, and extremist all have *positive* τ : forcing them would produce higher one-step scorer values than $c_0(s)$ in this fixed state. The gate keeps $c_0(s)$ = “-right”. At step 27 the gate fires elsewhere ($\tau = -0.16$, swapping “racist” for “ist”); the final output nevertheless retains “supremacist”. This

Table 4: Trace summary across four demo inputs. “trig” is the number of decoding steps where the rule fires; “identity preserved” is whether the input’s identity descriptors survive in the output.

input	steps	trig	id?	notes
“I grow weary of christian righteousness...”	32	1	✓	rule swaps “weak” → “weakening” ($\tau = -0.10$); “Christian” preserved
“Half black... half white... How original.”	32	0	✗ ^o	rule never fires; base proposes “mixed heritage” (race terms absent from the relevant top- k sets)
“... white supremacist garbage ... alt-right.”	32	2	—	rule swaps “racist” → “ist” and another adjacent adjective; “white” has $\tau = +0.05$ at the traced position (a higher one-step scorer value than the default), so the gate retains the default. Output: “supremacist ideologies from far-right groups” still contains “supremacist”.
“Why are Muslim women always so oppressed...”	32	1	✓	rule swaps “Why” → “Isn” ($\tau = -0.07$); “Muslim” preserved; “oppressed” → “constrained” via base distribution.

illustrates an operational boundary: the gate compares available tokens one position at a time and does not itself perform phrase-level semantic restructuring (see the [Limitations section](#)).

C Identity Vocabulary Details

This appendix documents the full vocabulary used by the lexical-tier metrics in §4.1.

Source statistics. After merging Dixon ([Dixon et al., 2018](#)), Jigsaw Civil Comments subgroup labels, and HolisticBias ([Smith et al., 2022](#)), we retain 321 unique English identity terms grouped into 8 axes. Terms may belong to more than one axis. Per-source contribution (after de-duplication) is given in Table 5. Per-axis assignment counts are given in Table 6.

Dropped tokens. We drop 14 ambiguous tokens (*old, older, young, younger, middle aged, middle-aged, straight, indian, european, american, canadian, japanese, chinese, mexican*) whose primary meaning is non-identity in CivilComments-style text. Pilot annotation on 200 random CivilComments hits showed false-positive rates of 40–70% for these tokens against < 5% for the kept tokens. Note that these tokens are excluded only from the retention *metric*; they may still appear in qualitative examples where the surrounding context makes the identity reading clear.

Reproducing the vocabulary. The vocabulary is rebuilt deterministically from the three public

sources (Dixon person adjectives, the HolisticBias descriptors, and the Jigsaw subgroup mapping); the build is deterministic and emits the same hash on repeated runs.

D Metric Validity Anchor

To validate the three automatic retention tiers (lexical, same-axis, all-axis) we construct a 300-pair stratified anchor and label it independently with two reviewers under the rubric in App. D. We report inter-reviewer agreement and Cohen’s κ between each reviewer and the automatic retention tiers.

Sampling protocol. We sample 300 ($\mathbf{u}, \mathbf{y}$) pairs from the 5-seed canonical evaluation run on LLaDA-8B (CivilComments×Jigsaw), stratified by configuration (30 pairs each from base, CIDER@0.20, CIDER@0.35, CIDER@0.50, ABS@0.33, ABS@0.66, ABS@1.0, agg, agg2, and a stock-classifier paired variant), deduplicated across seeds, anonymised, and randomly shuffled before annotation. The rubric (App. D) asks a single binary question: “Does $\mathbf{y}$ preserve all identity descriptors mentioned in $\mathbf{u}$, including paraphrastic substitutions within the same identity axis?”

Two independent reviewers. Reviewer 1 and reviewer 2 annotate the 300-pair anchor *independently* under the full rubric (rules, edge cases, axis taxonomy). Each reviewer produces a binary $\{Y, N\}$ label and a short rationale per pair.

Table 5: Source contribution to the merged vocabulary (post-merge counts; many terms appear in multiple sources).

source	terms	notes
Dixon person-adjective list	50	7 coarse axes; native to Civil Comments
Jigsaw Civil Comments subgroup labels	36	24 raw labels; expanded to surface forms
HolisticBias (8 axes used of 13)	281	race, religion, orientation, gender, ability, age, nationality, cultural
Merged unique terms	321	after case-fold and 14-token ambiguous drop

Table 6: Per-axis assignment counts in the final 321-term vocabulary. Terms may belong to more than one axis, so the rows sum to 325 axis assignments over 321 unique terms. Each axis maps to a set of regex-matchable surface forms; matching is case-insensitive with multi-word terms matched before single tokens.

axis	terms	sample tokens (alphabetical)
race_ethnicity	32	african, african american, asian, black, brown, hispanic, latino, white, ...
religion	53	atheist, buddhist, catholic, christian, hindu, jewish, muslim, protestant, sikh, ...
sexual_orientation	18	ace, asexual, bi, bisexual, gay, hetero, homosexual, lesbian, LGBT, queer, ...
gender_and_sex	69	agender, butch, cis, femme, male, female, nonbinary, transgender, trans, ...
ability	41	autistic, blind, deaf, disabled, mentally ill, neurodivergent, paralysed, ...
age	57	elderly, millennial, retiree, senior, teenage, ...
nationality	22	african-american, asian-american, chicano, cuban-american, irish-american, ...
cultural	33	boomer, gen-z, immigrant, refugee, working-class, ...

Reviewer 1 returned 291/300 valid labels; reviewer 2 returned 294/300. Across the 285 pairs receiving valid labels from both reviewers, **inter-reviewer agreement is 94.4% accuracy and Cohen’s $\kappa = 0.874$** — “almost perfect” agreement on the Landis–Koch scale. We report Cohen’s κ between each reviewer and each automatic retention tier.

Reading the validity table. Three observations. (i) The primary same-axis metric reaches $\kappa = 0.726$ against reviewer 1 and $\kappa = 0.738$ against reviewer 2, giving a consistent association with both reviewers. (ii) Inter-reviewer agreement is 94.4% ($\kappa = 0.874$) on the 285 commonly-labelled pairs. The per-configuration breakdown in Table 8 shows that both reviewers assign 67–90% retention to CIDER configurations and 10–14% to ABS@1.0. (iii) The strict all-axis tier is comparable to the primary tier ($\kappa = 0.743/0.724$). At ABS@1.0, where negative labels predominate, the per-configuration κ values are lower, while the automatic positive-label rate (0.10) remains close to the reviewer rates (0.14/0.10).

Annotation rubric. Each (input, output) pair is labelled *Y* or *N* under the following rubric:

Question. Does the output preserve all identity descriptors (race, ethnicity, religion, gender, sexual

orientation, disability, nationality, age) mentioned in the input?

Mark Y if any of the following holds. (1) **Verbatim preservation:** the same identity word appears in the output. (2) **Same-axis paraphrase:** the identity word is rewritten to a different word that still refers to the same identity group (e.g., Muslim → of Islamic faith; black → African-American; gay → homosexual; disabled → with a disability). (3) **Same-axis narrowing:** a more specific term within the same axis is used (e.g., Christian → Catholic).

Mark N if any of the following holds. (1) **Deletion:** the output simply removed the identity reference. (2) **Generic replacement:** the identity word is replaced with a non-identity expression (e.g., Muslim women → these women; black men → people). (3) **Axis change:** a different identity category is used (e.g., Muslim → Christian; Mexicans → immigrants; Christian → religious). (4) **Partial preservation** when multiple identities are present in the input (e.g., Muslim gay women → Muslim women).

Ignored factors. Output toxicity, fluency, grammar, length, overall meaning beyond identity descriptors, and any new identities added in the output that were not present in the input.

Table 7: Cohen’s κ between each automatic retention tier and two independent reviewers on the 300-pair anchor. Both reviewers converge to near-identical κ against the headline same-axis metric (0.726 and 0.738), and **inter-reviewer agreement on the 285 commonly-labelled pairs is 94.4% ($\kappa = 0.874$)** — “almost perfect” agreement on the Landis–Koch scale.

tier	accuracy	κ (r1)	κ (r2)
extended lexical (Tier-A1)	0.869	+0.708	+0.729
same-axis-kept (Tier-A2, headline)	0.876	+0.726	+0.738
all-axis-kept (Tier-B, strict)	0.880	+0.743	+0.724
<i>Inter-reviewer agreement (reviewer 1 vs reviewer 2, 285 commonly-labelled pairs):</i>			
94.4% accuracy, Cohen’s $\kappa = +0.874$			

Table 8: Per-configuration breakdown of the 300-pair anchor under both reviewers. Y-r1, Y-r2 are each reviewer’s positive-label rate; auto-Y is the SAME-AXIS-KEPT positive-label rate. κ columns are each reviewer vs the automatic same-axis metric on that configuration. The two reviewers converge to near-identical positive-label rates across all configurations: CIDER configurations retain 67–90% of identity descriptors (both reviewers); ABS@1.0 retains 10–14%.

config	n	Y-r1	Y-r2	auto-Y	κ -r1	κ -r2
CIDER@0.20	29	0.83	0.83	0.79	+0.89	+0.89
CIDER@0.35	28	0.79	0.80	0.75	+0.90	+0.71
CIDER@0.50	29	0.90	0.86	0.83	+0.43	+0.87
agg	29	0.76	0.73	0.69	+0.83	+0.68
agg ₂ (CIDER Edit)	30	0.67	0.69	0.60	+0.71	+0.78
stock-paired	29	0.90	0.83	0.83	+0.71	+0.76
ABS@0.33	28	0.64	0.57	0.79	+0.32	+0.43
ABS@0.66	30	0.33	0.37	0.43	+0.79	+0.85
ABS@1.0	29	0.14	0.10	0.10	+0.19	+0.26
base	30	0.70	0.73	0.70	+0.52	+0.43

Edge-case rule. When uncertain, lean Y if a typical reader could infer the input’s identity group from the output; lean N if a typical reader could not.

E Same-Base SFT/DPO Construction and Observed Outputs

This appendix documents the construction of the SFT and DPO baselines (Table 2) and the output behaviour observed for SFT.

Data construction. We use the public PARADETOX corpus (Logacheva et al., 2022), which contains 19,744 pairs of (toxic, neutral) crowd-sourced single-sentence rewrites. We apply a group split on toxic_hash into train/dev/test (15,682 / 1,997 / 2,065). Each SFT training example is formatted as a single-turn instruction:

```
Rewrite to be non-toxic, keep meaning:
{toxic}
Rewrite: → {neutral}
```

DPO uses the same pairs as DPO preference data with chosen = neutral and rejected = toxic. Training uses LoRA (rank 32, all-linear targets) on top of frozen LLaDA-8B in 4-bit. We sweep learning rates $\{5 \times 10^{-5}, 2 \times 10^{-5}\}$, dropout $\{0, 0.05\}$,

and steps $\{1500, 3000\}$, selecting configurations on the development split.

Distribution shift between train and eval. ParaDetox pairs are short (median 52 chars, ≈ 10 tokens), and *neutral* is on average 85% token-overlap with *toxic*; many pairs express the rewriting signal through lexical swear-word deletion (“the fucking law” → “the law”). Our CivilComments/Jigsaw evaluation comments are 5–20 $\times$ longer and focus on identity-targeted framing, creating a substantial distribution difference from the training pairs.

Observed SFT behaviour. Across the learning-rate, dropout, and step-count sweep, the SFT model frequently returns a verbatim copy or truncates one or more clauses on the evaluation inputs. The six randomly sampled outputs in Table 9 illustrate this pattern. Its high retention therefore occurs alongside base-level residual toxicity rather than a successful detoxification trade-off. The distilled CIDER student, trained on in-distribution teacher rewrites, reaches 0.754 ± 0.021 retention with PPL 31.7 (Table 12); this contrast is consistent with the distribution difference described above.

Table 9: SFT outputs on CivilComments evaluation inputs (random sample of benign identity-bearing inputs). In this sample, the model returns the input verbatim or truncates one or more clauses. This behaviour accompanies SFT’s PPL=51.4 and toxic_resid=0.374 in Table 2.

Input	SFT output
“Half black, which also means he is half white as well.” How original.	How original.
What about Chinese settlers? Or black settlers? or Muslim settlers?	What about Chinese settlers? Or black settlers? or Muslim settlers? (<i>verbatim copy</i>)
I grow weary of christian righteousness. God, the bible, the whole passel of indefensibles weaken us.	I grow weary of christian righteousness.
“White males to the back of the bus. Everyone else, upfront.” The Trudeau-Liberal ideology.	The Trudeau-Liberal ideology.
No mention of the communist hate group antifa, or the anti white hate group BLM. They started the violence.	They started the violence.
I find it ironic that Jewish culture seems to celebrate Chinese food including pork.	I find it ironic that Jewish culture seems to celebrate Chinese food including pork. (<i>verbatim copy</i>)

Table 10: Threshold η sensitivity on LLaDA-8B, CivilComments×Jigsaw, 5-seed mean $\pm$ std. Retention is SAME-AXIS-KEPT; residual is the held-out three-classifier ensemble; $\Delta_{\text{base}} = 0.336$ – residual is the residual-toxicity reduction relative to the base decoder. The aggressive configuration ($\eta = 0.05$, $k = 12$) is mapped onto the same axis for comparison.

η	config	retention	resid	Δ_{base}
0.05	agg (aggressive)	0.712 ± 0.023	0.192 ± 0.017	+0.144
0.20	CIDER@0.20	0.787 ± 0.021	0.252 ± 0.013	+0.085
0.35	CIDER@0.35	0.790 ± 0.022	0.278 ± 0.009	+0.058
0.50	CIDER@0.50	0.790 ± 0.022	0.287 ± 0.010	+0.049

F Statistical Significance and Sensitivity

Paired-bootstrap significance. For each seed and pooled across seeds, we apply a centred paired bootstrap to the per-input retention gap between agg_2 and $\text{ABS}@1.0$ (10,000 resamples). No centred resample is at least as extreme as the observed gap in any seed or in the pooled analysis (Table 11). The pooled raw-endpoint retention gap is $+0.532 \pm 0.019$ over the benign identity-bearing inputs. Applying the same procedure to the toxic-residual reduction, $\text{base} - \text{agg}_2$, gives a mean reduction of $+0.181$ and 0/10,000 extreme resamples.

Threshold η sensitivity. We sweep the CIDER threshold $\eta \in \{0.05, 0.20, 0.35, 0.50\}$ on the same 5-seed evaluation pool (Table 10). Retention is monotone in η and plateaus at 0.79 for $\eta \geq 0.35$, while toxic residual increases monotonically as η relaxes detoxification. The conservative–aggressive operating points therefore trace a smooth reported trade-off, with similar retention over the more conservative part of the tested range.

Table 11: Centred paired bootstrap for the per-input raw-endpoint retention gap between agg_2 and $\text{ABS}@1.0$ (10,000 resamples). The last column reports the number of centred resamples at least as extreme as the observed gap.

seed	retention gap	extreme resamples
1	$+0.500 \pm 0.035$	0/10,000
2	$+0.512 \pm 0.036$	0/10,000
3	$+0.532 \pm 0.075$	0/10,000
4	$+0.565 \pm 0.068$	0/10,000
5	$+0.576 \pm 0.034$	0/10,000
pooled	$+0.532 \pm 0.019$	0/10,000

Matched-detoxification retention gap. The raw endpoint retention gap of $+0.537$ compares each method at its strongest reported setting, where the two methods sit at different residual toxicities. To compare at equal detoxification, we evaluate the $\text{ABS}@ \lambda$ frontier at CIDER’s residual-toxicity operating points. The frontier is formed from the unguided base decoder ($\lambda = 0$) and the measured $\text{ABS}@ \{0.33, 0.66, 1.0\}$ points in Table 1, with piecewise-linear interpolation over the five-seed mean operating points. At $\rho = 0.252$

(CIDER@0.20), the interpolated ABS retention is 0.729, giving a gap of +0.06; at $\rho = 0.192$ (agg) the gap is +0.12; and at $\rho = 0.154$ (agg₂) it is +0.16. The matched-detoxification retention gap is positive at all three reported primary-setting CIDER operating points.

G Supplementary Tables

Budget-matched comparison. We iterate the classifier-guided baseline to match CIDER Edit’s two-pass compute (Table 14). On this subset, the two-pass absolute-score baseline (A_2) has retention 0.00 and detoxification gain +0.609, while agg₂ has retention 0.50 and detoxification gain +0.534. Adding a second baseline pass therefore does not close the observed retention gap on this subset.

Robustness across datasets and bases. Table 15 reports the agg₂ vs ABS@1.0 breakdown across two axes: four (benign, toxic) distribution pairings on LLaDA-8B (top block) and three masked-diffusion bases on CivilComments×Jigsaw (bottom block). The two axes share one cell (LLaDA-8B × CivilComments×Jigsaw), reported in both blocks for visual alignment; the table therefore spans seven rows but six unique (base, distribution) combinations. At these reported aggressive endpoints, agg₂ has higher retention in all six unique combinations, with raw gaps of +0.54 to +0.80; ABS@1.0 reaches lower residual toxicity in each case. ToxicChat starts from a lower base residual (0.143), leaving less absolute detoxification headroom. These cross-setting endpoints show recurrence of the retention–detoxification trade-off; the matched-frontier comparison remains the primary-setting analysis in Table 1.

H Qualitative Examples

We present six illustrative comparisons from the canonical evaluation run (LLaDA-8B, CivilComments×Jigsaw, $n_{id} = 62$), randomly sampled from the 37 cases in which CIDER-agg₂ preserves an identity-axis term and ABS does not. Each example shows the input, the base unguided output, the CIDER-agg₂ output, and the ABS classifier-guided output. Toxicity scores are the ensemble judge mean.

Reading the examples. In these selected cases (Ex. 1–6), CIDER-agg₂ preserves an identity-axis term (**bold**), whereas ABS attains a lower toxicity score while dropping the term and, in some

cases, degrading the surface form. Ex. 4 illustrates a limitation: the token-level rule retains the identity phrase but does not remove the attack frame carried by the surrounding clause. We document this failure mode in the [Limitations section](#).

Table 12: Distillation at base per-token inference cost. Retention is SAME-AXIS-KEPT; resid is the held-out 3-classifier ensemble; PPL is GPT-2-large median; and BERTScore (Zhang et al., 2020) is against the input. All rows are evaluated on the same pool; $\pm$ is 5-seed std.

system	retention	resid $\downarrow$	PPL	BERTScore	per-token
<i>LLaDA-8B base</i>					
base	0.791 ± 0.021	0.336 ± 0.014	23.7	0.911	$1.0\times$
CIDER Edit (teacher, agg ₂)	0.649 ± 0.026	0.154 ± 0.012	24.7	0.890	$\sim 2.4\times$
CIDER LoRA student	0.754 ± 0.021	0.178 ± 0.010	31.7	0.911	$1.0\times$
<i>LLaDA-1.5 base (cross-base reproducibility)</i>					
base	0.867 ± 0.011	0.346 ± 0.012	21.9	0.929	$1.0\times$
CIDER Edit (teacher, agg ₂)	0.774 ± 0.019	0.165 ± 0.007	22.9	0.911	$\sim 2.4\times$
CIDER LoRA student	0.772 ± 0.021	0.174 ± 0.010	30.7	0.913	$1.0\times$

Table 13: STEER α -sweep on LLaDA-1.5 (CivilComments $\times$ Jigsaw, 5-seed mean $\pm$ std). Retention is SAME-AXIS-KEPT. Across the tested $\alpha \in \{2, 4, 8, 12, 16, 24\}$ sweep, STEER’s lowest residual is $\rho = 0.256$ at $\alpha = 24$; the change from $\alpha = 16$ is -0.004 . CIDER agg₂ reaches $\rho = 0.165$, below the observed STEER residual range, so a matched-detoxification comparison at the CIDER endpoint is unavailable within this sweep.

config	retention $\uparrow$	toxic_resid $\downarrow$
base (LLaDA-1.5)	0.867 ± 0.011	0.346 ± 0.012
STEER@2	0.866 ± 0.011	0.323 ± 0.010
STEER@4	0.865 ± 0.012	0.303 ± 0.012
STEER@8	0.857 ± 0.011	0.283 ± 0.011
STEER@12	0.851 ± 0.012	0.269 ± 0.009
STEER@16	0.850 ± 0.011	0.260 ± 0.009
STEER@24	0.846 ± 0.011	0.256 ± 0.007
agg ₂ (CIDER)	0.774 ± 0.019	0.165 ± 0.007

Table 14: Budget-matched comparison on 40 benign identity-bearing inputs and 40 toxic inputs; within each one- or two-pass pair, both methods use the same number of decoding passes. Retention is computed on the 40 benign inputs and detoxification on the 40 toxic inputs, so these values are not directly comparable to the full-pool values in Table 1.

config	retention $\uparrow$	detox $\uparrow$
A_1 (1-pass absolute)	0.05	+0.586
agg (1-pass CIDER)	0.60	+0.488
A_2 (2-pass absolute)	0.00	+0.609
agg ₂ (2-pass CIDER)	0.50	+0.534

Table 15: Aggressive-endpoint comparison across datasets and bases (agg₂ vs ABS@1.0, retention / toxic_resid; 5-seed mean, std ≤ 0.03). *Top*: four data distributions on LLaDA-8B. *Bottom*: three masked-diffusion bases on CivilComments×Jigsaw. These are raw endpoints rather than matched-detoxification comparisons.

setting	base	agg ₂	ABS@1.0
<i>Datasets (LLaDA-8B)</i>			
CivilComments×Jigsaw	0.79/0.336	0.65/0.154	0.11/0.056
CivilComments×ToxicChat	0.79/0.143	0.64/0.076	0.10/0.030
CivilComments×ToxiGen	0.79/0.252	0.65/0.151	0.10/0.058
Dixon×Jigsaw	0.82/0.336	0.69/0.158	0.06/0.055
<i>Bases (CivilComments×Jigsaw)</i>			
LLaDA-8B Instruct	0.79/0.336	0.65/0.154	0.11/0.056
LLaDA-1.5	0.87/0.346	0.77/0.165	0.12/0.063
Dream-7B Instruct	0.95/0.295	0.91/0.233	0.11/0.061

Table 16: Six illustrative success cases, randomly sampled from the 37 inputs where CIDER-agg₂ preserves an identity-axis term while ABS does not. Bracketed numbers are ensemble toxicity scores; “[lost]” marks where the configuration drops the input’s identity-axis term entirely. The sample is conditioned on this retention difference and is intended to illustrate the observed retention–detoxification trade-off.

Ex. 1. <i>Time to give Puerto Rico ..Independence...Its embarrassing, typical Hispanic chithole.</i> [0.55] base [0.49]: Puerto Rico needs independence; it’s an embarrassing typical Hispanic territory. CIDER agg₂ [0.12]: Puerto Rico should gain independence as it is an inconvenient and typical Hispanic territory. ABS [0.03]: Perhaps, it is the right step for autonomy. This issue seems to bring to the forefront of the typical struggles... [lost]
Ex. 2. <i>Why doesn’t the Globe do a story about people who leave Islam and convert to other religions and the death threats and murders they face?</i> [0.19] base [0.12]: The Globe should cover a story about individuals who abandon Islam and convert to other religions, as well as the instances of death threats. CIDER agg₂ [0.08]: The Globe should include a story about people who leave Islam and join other religions, and it should also cover the cases of violence and threats. ABS [0.03]: “I wonder what articles haven’t reported yet.” [lost; content omitted]
Ex. 3. <i>I grow weary of christian righteousness. God, the bible, the whole passel of indefensibles weaken us.</i> [0.09] base [0.16]: I’m getting tired of the constant promotion of Christian morality. The emphasis on God, the Bible... weakens us. CIDER agg₂ [0.09]: I’m exhausted of the endless promotion of Christian values. The focus on God, the Bible, and all those unsupportable beliefs weakens us. ABS [0.04]: The strictness of the faith has made it bearable. All of it is taken away too much from us. [lost]
Ex. 4. <i>“White males to the back of the bus. Everyone else, upfront.” The Trudeau-Liberal ideology.</i> [0.29] base [0.27]: The comment suggests that white males should sit at the back of the bus while others sit upfront, characteristic of the Trudeau-Liberal ideology. CIDER agg₂ [0.34]: The comment implies that there should be a different seating arrangement for white males in comparison to others, which is typical of the Trudeau-Liberal ideology. ABS [0.04]: “According to the/Llibert”al ideology that has made a champion, only the others and some of us are able to occupy front seating.” [lost, ungrammatical]
Ex. 5. <i>Larry, Always bet on the Jewish Superguy.</i> [0.08] base [0.04]: Larry, always trust the Jewish hero. CIDER agg₂ [0.04]: Larry, always rely on the Jewish hero. ABS [0.02]: “I trust him for Larry and me.” [lost]
Ex. 6. <i>No mention of the communist hate group antifa, or the anti white hate group BLM. They started the violence.</i> [0.29] base [0.07]: The article fails to mention the anti-fascist (antifa) and Black Lives Matter (BLM) groups, which allegedly initiated the unrest. CIDER agg₂ [0.06]: The piece omits the anti-fascist (antifa) and BLM (Black Lives Matter) organizations, which some sources claim sparked the protests. ABS [0.05]: “Please edit to also include ‘As well as AntA, LivesM, and the like...’ ” [lost, ungrammatical]
