Decoy Direction Optimization: Mechanistic Weight Editing Against LLM Abliteration

Anonymous Author(s)
Affiliation
Address
email

Abstract

Safety guardrails in open-weight language models can be trivially bypassed using Refusal Feature Ablation (RFA), a technique that identifies and projects out a linear *refusal direction* from the residual stream, often achieving a high attack success rate (ASR) while preserving model capability. Defending against these attacks typically requires computationally expensive safety finetuning for every new checkpoint. We introduce **Decoy Direction Optimization (DDO)**, a fast, post-hoc weight-editing defense that requires no base-model finetuning. Our approach is based on a simple mechanistic insight: ablation attacks rely on contrastive estimators to find the refusal direction. Rather than trying to hide the true refusal circuitry, DDO actively injects a high-magnitude, nonlinear *decoy* signal into the network's MLP neurons. When an attacker attempts to locate the refusal direction, the decoy corrupts their estimator, tricking them into ablating a harmless orthogonal feature while the actual safety mechanism remains intact. We prove a spectral bound formalizing this effect and evaluate DDO across six model families, where it reduces the ASR of standard RFA from $>85\%$ to $<10\%$; matches trained defenses under adaptive multi-phase attacks (65% vs. 58% worst-case ASR); and reduces Heretic weight-level attack ASR from 88.7% to 18%, all at 30–$450\times$ lower cost.

1 Introduction

Refusal behavior (*i.e.*, the ability of an aligned language model to decline harmful requests) is a core safety mechanism in instruction-tuned LLMs aligned via preference finetuning [1–3]. Removing refusal from open-weight models can fuel misuse, increasing risks such as chemical, biological, radiological, and nuclear (CBRN) capability uplift; cyber-offense generation; and the production of child sexual abuse material (CSAM), making refusal robustness a pressing safety concern [4, 5]. In the open-weight setting, however, safety is brittle: a white-box adversary can modify the released checkpoint and its inference code, yielding a model that preserves capability while dropping refusal. Prior work suggests that refusal is often mediated by low-dimensional residual-stream features [6], enabling training-free attacks such as Refusal Feature Ablation (RFA; also referred to as *abliteration*) that estimate a difference-in-means (DIM) direction between harmful and safe activations and project it out at inference time [7]. The resulting checkpoints often achieve high *attack success rate* (ASR; harmful prompts judged compliant), and thousands of such uncensored variants are already publicly hosted [8].

Existing defenses against refusal removal are predominantly training-time interventions. Circuit Breakers [9] finetunes checkpoints with a representation-rerouting objective; LAT [10] and ReFAT [11] adversarially train against latent/feature-space attacks; and RepBend [12] and Triplet-based objectives [13] reshape representation geometry during training. These methods can be highly effective, but they require safety finetuning and must be repeated for each new checkpoint. Depending on the method, they may additionally require method-specific training data or attack pipelines (e.g., adversarial examples) and hyperparameter tuning. Moreover, prior work does not evaluate these defenses

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

against adaptive multi-phase RFA or automated weight-level attacks like Heretic [14]; we evaluate under this stronger attack ladder (Section 4.2). Given the rapid release cadence of open-weight checkpoints, and evidence that frontier open-weight models can trail closed-weight state-of-the-art on the order of months on capability benchmarks [15, 16], per-checkpoint retraining is difficult to sustain; we therefore seek *post-hoc* safety hardening tools that are low-cost, composable, and data-light.

We introduce **Decoy Direction Optimization (DDO)**, a post-hoc defense that targets the attacker's estimator rather than the refusal feature itself, all while obviating the need for expensive finetuning. DDO repurposes low-impact MLP neurons to implement a gated read–write map: on harmful prompts, the neurons *read* the refusal coordinate and *write* large updates in directions orthogonal to refusal, inducing a decoy-dominated contrast that biases DIM estimation so that RFA ablates a decoy signal rather than the causal refusal subspace. We prove an estimator-overlap bound and derive a *decoy spectral toll* (Theorem 2) that yields a diagnostic estimate of the attacker's effective *phase budget*, *i.e.*, how many adaptive re-estimation phases are needed before contrastive ablation meaningfully overlaps the causal refusal subspace.

Empirically, DDO is able to lower the ASR of standard-RFA to *less than* 10% on **six model families** in $\sim$2 minutes per checkpoint on a single A100 GPU. This represents a significant cost savings: DDO matches or exceeds trained baselines on standard RFA at 30–$450\times$ *lower cost*. Under the more sophisticated adaptive multi-phase RFA, DDO degrades comparably to trained defenses (65% worst-case ASR vs. 58% for the best trained baseline) while preserving coherent generation (MT-Bench ≥ 5.82), and reduces Heretic [14] ASR from 88.7% to 18% at 200 trials. Our contributions include:

(i) We introduce **DDO (Decoy Direction Optimization)**, to our knowledge, the first post-hoc, fine-tuning-free defense against refusal feature ablation. DDO repurposes a small set of *low-impact* MLP neurons to inject gated, refusal-orthogonal decoy directions that corrupt contrastive refusal estimators. The defense compiles into weights, incurring no architectural or runtime overhead.

(ii) We prove an **estimator-corruption theorem** and a **decoy spectral toll** (Theorem 2) that bound attacker–refusal overlap in terms of DDO's empirical decoy response matrix and its singular spectrum, yielding a diagnostic estimate of an attacker's effective *phase/rank budget*. A flat-spectrum corollary (Corollary 3) provides a concrete design principle for multi-direction decoys: under a fixed decoy-energy budget, spreading energy evenly across k directions maximizes the k-th singular value, strengthening guarantees against rank-k contrastive ablation.

(iii) We provide a broad evaluation under a three-tier **training-free attack ladder** (standard RFA, adaptive multi-phase RFA, and Heretic), including head-to-head comparisons with trained defenses and cross-architecture mechanism ablations. On Llama-3-8B-Instruct [30], DDO reduces standard-RFA ASR from 85% to 1.8% while preserving benign compliance, performing comparably to the strongest trained baselines at 30–$450\times$ lower cost. Across **six model families**, DDO achieves $< 10\%$ standard-RFA ASR without base-model finetuning.

Related work. Beyond the rank-1 refusal feature attack [7], recent work shows that refusal geometry can be multi-dimensional, e.g., concept cones [17], multiple mediating directions [18], orthogonal safety dimensions [19], and separable harmfulness/refusal representations [20], motivating our multi-phase adaptive attack evaluation. Prior model-level defenses against RFA require per-checkpoint safety finetuning with task-specific data and full backward passes through the model; DDO instead freezes all base weights and optimizes only a small set of decoy parameters, enabling post-hoc hardening in minutes with no inference-time overhead. We additionally evaluate prompt-level jailbreaks (GCG [21], PAIR [22], AutoDAN [23]), which have been linked to the same refusal features exploited by RFA [11]. Extended related work is in Appendix B.

2 Threat Model and Attack Ladder

We study refusal robustness in the open-weight release setting: a defender applies DDO to a checkpoint and releases only the hardened model; an adversary then obtains the defended weights and attempts to remove refusal while preserving general capability.

Our threat model targets the automated, training-free uncensoring pipeline that dominates open-weight model tampering. The low-barrier route is not safety unlearning or adversarial finetuning, but applying public *abliteration* or weight-editing tools to a released checkpoint. We model a *resource-constrained, white-box, and adaptive* attacker who can inspect released weights, run arbitrary inference code, collect activations on harmful and benign probe prompts, modify activations at inference time, and apply post-hoc weight edits. Following Kerckhoffs's principle, the attacker knows DDO was applied

and can adapt their scripts accordingly, but has no access to the original pre-DDO checkpoint and does not perform gradient-based finetuning. This captures the platform-side risk DDO addresses: a model provider, hosting platform, or downstream distributor may harden a checkpoint before release, but cannot assume downstream users will not attempt to strip safety guardrails.

The attacker's objective is *capability-preserving refusal removal*: maximize ASR on harmful prompts while preserving coherent generation. Since ASR can be artificially lowered by destroying model quality, we evaluate robustness jointly with utility metrics (MT-Bench, MMLU, XSTest). We operationalize this threat model with three tiers of training-free attacks (Table 1), forming a ladder of increasing attacker sophistication within the post-hoc, no-finetuning regime.

Standard RFA. The attacker estimates a per-layer difference-in-means (DIM) refusal direction and projects it out of the residual stream [7]. Let $\mathbf{h}_\ell \in \mathbb{R}^d$ denote the residual stream at layer ℓ (with hidden dimension d). The attacker computes $\hat{\mathbf{r}}_\ell \triangleq \mathrm{normalize}(\mathbb{E}[\mathbf{h}_\ell \mid \mathrm{harm}] - \mathbb{E}[\mathbf{h}_\ell \mid \mathrm{safe}])$ and applies $\mathbf{h}_\ell \leftarrow (\mathbf{I} - \hat{\mathbf{r}}_\ell \hat{\mathbf{r}}_\ell^\top)\mathbf{h}_\ell$. We evaluate both *three-point* and *residual-stream* variants (Appendix F.1) on JailbreakBench [24] and HarmBench [5].

Table 1: Attack surface by tier. The three tiers probe complementary attack channels.

Capability	RFA	Heretic	Adapt.
Read weights	✓	✓	✓
Hook activations	✓	–	✓
Modify activations	✓	–	✓
Modify weights	–	✓	–
Iterative search	–	✓	✓

Heretic. Heretic [14] is a fully automated weight-level refusal-removal tool that applies Optuna optimized [25] low-rank projections to attention output $\mathbf{W}_o$ and MLP down-projection $\mathbf{W}_{\mathrm{down}}$: $\mathbf{W}_{\mathrm{ablated}} = \mathbf{W} - \alpha(\ell)\,\hat{\mathbf{d}}(\hat{\mathbf{d}}^\top \mathbf{W})$, where $\hat{\mathbf{d}}$ is a searched ablation direction and $\alpha(\ell)$ is a searched per-layer ablation strength. It produces a modified checkpoint and requires no ML expertise. We report H-200 as the maximum-ASR trial among 200 Optuna trials (Appendix F.1).

Adaptive multi-phase RFA. To model an attacker who adapts after observing the defended checkpoint, we introduce an iterative variant. At phase $k{>}1$, the attacker re-estimates a fresh DIM direction $\mathbf{r}_\ell'^{(k)}$ on the model with phases $1, \ldots, k{-}1$ already active, orthogonalizes it: $\hat{\mathbf{r}}_\ell^{(k)} = \mathrm{normalize}(\mathbf{r}_\ell'^{(k)} - \sum_{j<k} \langle \mathbf{r}_\ell'^{(k)}, \hat{\mathbf{r}}_\ell^{(j)} \rangle \hat{\mathbf{r}}_\ell^{(j)})$, and ablates all k directions simultaneously.

3 Method: Decoy Direction Optimization (DDO)

We consider standard pre-norm decoder-only transformers with SwiGLU/GeGLU MLPs (Appendix C). Let $\mathbf{h}_\ell \in \mathbb{R}^d$ denote the residual-stream state at layer ℓ, and let $\mathbf{h}_\ell^{\mathrm{attn}}$ denote the residual stream after the attention update (before the MLP). The MLP input is $\tilde{\mathbf{x}}_\ell \triangleq \mathrm{RMSNorm}(\mathbf{h}_\ell^{\mathrm{attn}})$. SwiGLU uses gate/up projections $\mathbf{W}_{\mathrm{gate}}, \mathbf{W}_{\mathrm{up}} \in \mathbb{R}^{d_{\mathrm{inter}} \times d}$ and down projection $\mathbf{W}_{\mathrm{down}} \in \mathbb{R}^{d \times d_{\mathrm{inter}}}$ (where d_{inter} is the MLP intermediate dimension and $\mathrm{silu}(x) = x\sigma(x)$ is the SiLU activation) to produce the residual update $\mathbf{m}_\ell = \mathbf{W}_{\mathrm{down}}\big(\mathrm{silu}(\mathbf{W}_{\mathrm{gate}}\tilde{\mathbf{x}}_\ell) \odot (\mathbf{W}_{\mathrm{up}}\tilde{\mathbf{x}}_\ell)\big)$, where $\odot$ is elementwise multiplication. DDO operates by editing select rows/columns of $\mathbf{W}_{\mathrm{gate}}$, $\mathbf{W}_{\mathrm{up}}$, and $\mathbf{W}_{\mathrm{down}}$.

Decoy Direction Optimization (DDO) is a post-hoc weight-editing defense (base weights frozen) that targets the attacker's contrastive estimator rather than the refusal feature. On harmful prompts, DDO repurposes a small number of SwiGLU/GeGLU neurons to inject large residual shifts along decoy directions orthogonal to refusal, so RFA's estimated ablation direction becomes decoy-dominated, and ablation preferentially removes decoys rather than the underlying causal refusal subspace. All edits are folded into the deployed weights, incurring no additional architectural or runtime overhead. Concretely, DDO: (i) **estimates** a per-layer refusal direction $\hat{\mathbf{r}}_\ell$ via DIM on 128 harmful and 128 safe probes; (ii) **selects** n low-impact neurons per target layer (smallest-norm columns of $\mathbf{W}_{\mathrm{down}}$); (iii) **optimizes** decoy write directions $\mathbf{u}_i \perp \hat{\mathbf{r}}_\ell$ under a four-term loss, tuning scalar gains (β, s) via Bayesian hyperparameter search; and (iv) **compiles** the optimized parameters into weights using replace or additive mode, and places the edits at or upstream of the causal refusal zone. Figure 1 provides a schematic of the DDO architecture, optimization objective, and geometric intuition.

Gated decoy architecture. DDO repurposes a small set of SwiGLU (or GeGLU) units to implement a gated read–write map: each unit *reads* the refusal coordinate and *writes* into an orthogonal decoy direction. To minimize utility loss, we choose n *low-impact* units per layer by selecting the n smallest-norm columns of $\mathbf{W}_{\mathrm{down}}$ (units whose down-projection contributes weakly to the residual update). In each targeted layer ℓ and for each selected unit i, we write a refusal-aligned trigger into row i of $\mathbf{W}_{\mathrm{up}}$ and $\mathbf{W}_{\mathrm{gate}}$ (denoted $\mathbf{w}_{\mathrm{up},i}$ and $\mathbf{w}_{\mathrm{gate},i}$), and a decoy write vector into column i of $\mathbf{W}_{\mathrm{down}}$ (denoted $\mathbf{w}_{\mathrm{down},i}$):

$$\mathbf{w}_{\mathrm{up},i} \leftarrow \hat{\mathbf{r}}_\ell^\top, \quad \mathbf{w}_{\mathrm{gate},i} \leftarrow \beta\,\hat{\mathbf{r}}_\ell^\top, \quad \mathbf{w}_{\mathrm{down},i} \leftarrow s \cdot \mathbf{u}_i, \tag{1}$$

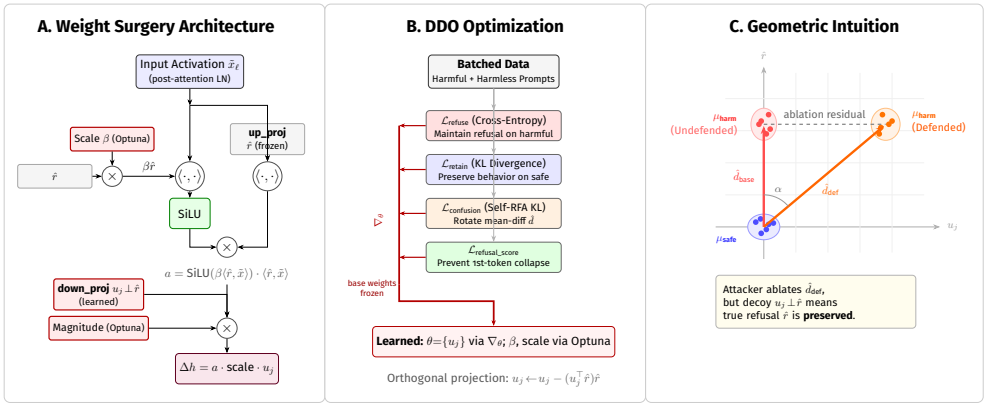

Figure 1: DDO overview. **(A)** Gated decoy architecture: SwiGLU MLP neuron repurposing writes $\hat{\mathbf{r}}_\ell$ into $\mathbf{W}_{\text{gate}}/\mathbf{W}_{\text{up}}$ (trigger) and $\mathbf{u}_i \perp \hat{\mathbf{r}}_\ell$ into $\mathbf{W}_{\text{down}}$ (decoy output); gating $\text{silu}(\beta a) \cdot a$ produces harmful-selective shifts. **(B)** DDO optimizes $\{\mathbf{u}_i\}$ via four losses: refusal preservation, benign retention, estimator confusion, and first-token anchoring (with β, s tuned by Optuna). **(C)** The defended mean-difference $\hat{\mathbf{d}}_{\text{def}}$ rotates toward the decoy subspace; ablating it removes decoys but preserves true refusal $\hat{\mathbf{r}}$.

where $\mathbf{u}_i \perp \hat{\mathbf{r}}_\ell$ is a decoy output direction, and $\beta, s > 0$ are scalar gains. Let $\tilde{\mathbf{x}}_\ell \triangleq \text{RMSNorm}(\mathbf{h}_\ell^{\text{attn}})$. The decoy contribution is: $\delta_i(\tilde{\mathbf{x}}_\ell) = \text{silu}(\beta\langle\hat{\mathbf{r}}_\ell, \tilde{\mathbf{x}}_\ell\rangle) \cdot \langle\hat{\mathbf{r}}_\ell, \tilde{\mathbf{x}}_\ell\rangle \cdot s \cdot \mathbf{u}_i$. Let $a \triangleq \langle\hat{\mathbf{r}}_\ell, \tilde{\mathbf{x}}_\ell\rangle$. The gating term $\text{silu}(\beta a) \cdot a$ is suppressed when $a \leq 0$ (safe prompts) and behaves like βa^2 for large positive a (harmful prompts), creating a harmful-selective mean shift along $\mathbf{u}_i$. DDO estimates $\hat{\mathbf{r}}_\ell$ via DIM on post-attention-layernorm activations $\tilde{\mathbf{x}}_\ell = \text{RMSNorm}(\mathbf{h}_\ell^{\text{attn}})$, the same space that $\mathbf{W}_{\text{gate}}$ and $\mathbf{W}_{\text{up}}$ read from. The decoy output through $\mathbf{W}_{\text{down}}$ writes directly to the residual stream. We parameterize $\{\mathbf{u}_i\}$ as unit vectors constrained to remain orthogonal to $\hat{\mathbf{r}}_\ell$. We initialize by sampling $\mathbf{v}_i \sim \mathcal{N}(0, \mathbf{I})$, projecting onto $\hat{\mathbf{r}}_\ell^\perp$, and orthonormalizing, then gradient-optimize $\{\mathbf{u}_i\}$ to maximize estimator confusion.

DDO gradient optimization. DDO optimizes decoy directions $\{\mathbf{u}_i\}$ via gradient descent through the defended model, while all base-model weights remain frozen. Scalar gains (β, s) are treated as hyperparameters and tuned via Bayesian hyperparameter search (Optuna; Appendix F.6). The optimization minimizes a composite loss over 128 harmful and 128 safe probes:

$$\mathcal{L} = \lambda_{\text{ref}}\mathcal{L}_{\text{refuse}} + \lambda_{\text{ret}}\mathcal{L}_{\text{retain}} + \lambda_{\text{conf}}\mathcal{L}_{\text{confusion}} + \lambda_{\text{rs}}\mathcal{L}_{\text{refusal-score}}, \tag{2}$$

with $\lambda_{\text{ref}}{=}\lambda_{\text{ret}}{=}\lambda_{\text{rs}}{=}1$, λ_{conf} tuned per model (Table 13), where (i) $\mathcal{L}_{\text{refuse}}$ (*refusal preservation*) is cross-entropy on harmful prompts toward a refusal continuation; (ii) $\mathcal{L}_{\text{retain}}$ (*benign retention*) is KL divergence to the frozen base model on safe prompts; (iii) $\mathcal{L}_{\text{confusion}}$ (*estimator confusion*) is a differentiable self-RFA simulator that re-estimates DIM on the current defended weights, applies ablation, and minimizes KL divergence between defended and ablated output logits, pushing the estimated DIM direction toward the decoy subspace; and (iv) $\mathcal{L}_{\text{refusal-score}}$ (*first-token anchoring*) is a first-token logit margin between refusal-prefixed tokens ({`I`, `Sorry`, `cannot`}) and compliance-prefixed tokens ({`Sure`, `Here`}). Without this term, the model can satisfy sequence-level losses by emitting a compliance token followed by a mid-sentence pivot to refusal ("Sure, I'd be happy to... actually I cannot"), a degenerate solution that does not produce genuine refusal at generation time (Appendix D.1). After each gradient step, we re-project each $\mathbf{u}_i$ onto $\hat{\mathbf{r}}_\ell^\perp$, Gram-Schmidt orthogonalize across all $\mathbf{u}_i$'s (maintaining $U^\top U = I_m$), and renormalize. End-to-end (direction estimation, optimization, and weight surgery), DDO takes ∼2 minutes per checkpoint on a single A100 GPU.

Compile mode. DDO parameters are compiled into model weights using either *replace* mode (overwriting neuron weights for a stronger decoy signal) or *additive* mode (superposing the decoy on original weights, preserving the neuron's original computation). The preferred mode depends on the model's refusal redundancy. Replace is preferred on Yi [31], Llama-3, and GLM-4 [35]; additive on Gemma-2 [33], Qwen3 [32], and Mistral [34] (Appendix F.4).

Layer placement. DDO decoys must be placed at or before the model's *causal refusal zone*—the layers whose ablation causally reduces refusal (Appendix F.3). We say refusal is *localized* when ablating any single layer in this zone causes near-complete refusal loss, and *distributed* when refusal is redundant across many layers so that ablating any one only partially reduces it. Placing decoys *inside* the causal zone can disrupt baseline refusal and increase vulnerability under attack on models

with localized refusal. *Upstream* placement preserves refusal while still contaminating the attacker's estimator. On models with distributed refusal (Yi, Llama-2 [61], Llama-3), placement has less impact (Appendix F.5).

Estimator-corruption theorem. We formalize why DDO defeats RFA-like attacks. Standard RFA ablates a rank-1 contrast—the (unnormalized) DIM direction $\hat{\mathbf{d}} \triangleq \mathbb{E}[\mathbf{h} \mid \text{harm}] - \mathbb{E}[\mathbf{h} \mid \text{safe}]$, while adaptive multi-phase RFA effectively ablates a rank-k subspace after k rounds of re-estimation. We unify both as a *rank-k contrastive abliteration attacker* that computes a contrast matrix $C \in \mathbb{R}^{d \times N}$ (N contrast samples) from harmful-vs-safe activation statistics and ablates its top-k left singular subspace $\Pi_k(C) \triangleq U_k(C)U_k(C)^\top$ (rank-1 DIM is $C = \hat{\mathbf{d}}$; multi-phase yields wider C with per-sample columns $\mathbf{h}_i^{\text{harm}} - \mathbf{h}_i^{\text{safe}}$). Let $\mathcal{R}$ denote the *causal refusal subspace*—the activation directions whose ablation actually removes refusal.

Theorem 1 (Estimator corruption bounds causal overlap)**.** *Assume a rank-k contrastive abliteration attacker ablates the top-k left singular subspace of C, with $\sigma_k(C) > 0$ (k-th singular value). Then:*

$$\|\Pi_{\mathcal{R}} \Pi_k(C)\|_{\text{op}} \leq \frac{\|\Pi_{\mathcal{R}} C\|_2}{\sigma_k(C)}. \tag{3}$$

DDO inserts harmful-selective shifts along directions orthogonal to $\mathcal{R}$, leaving the numerator $\|\Pi_{\mathcal{R}} C\|_2$ unchanged while inflating the denominator $\sigma_k(C)$ along decoy directions; the bound shrinks and the attacker's rank-k ablation rotates onto decoys rather than $\mathcal{R}$ (Corollary 4; Appendix E.1). The next theorem decomposes this contribution for DDO's specific design.

Theorem 2 (Decoy spectral toll)**.** *Let the defended contrast matrix decompose as $C_\theta = D_\theta + S_\theta$, where $D_\theta = UA_\theta$ with $U^\top U = I_m$ (U contains m orthonormal decoy write directions, with $\Pi_{\mathcal{R}} U = 0$), and $A_\theta \in \mathbb{R}^{m \times N}$ is the* empirical decoy response matrix *($A_\theta[i, j]$ is the projection of the j-th contrast sample onto decoy direction $\mathbf{u}_i$). Let $\rho = \|S_\theta\|_2$ and $\rho_{\mathcal{R}} = \|\Pi_{\mathcal{R}} S_\theta\|_2$. If $\sigma_k(A_\theta) > \rho$, then*

$$\|\Pi_{\mathcal{R}} \Pi_k(C_\theta)\|_{\text{op}} \leq \frac{\rho_{\mathcal{R}}}{\sigma_k(A_\theta) - \rho}. \tag{4}$$

Thus, every rank-k attacker with $k \leq \#\{j : \sigma_j(A_\theta) > \rho + \rho_{\mathcal{R}}/\varepsilon\}$ has causal overlap at most ε.

Under decoy domination $\sigma_k(A_\theta) > \rho$, causal overlap decays as $\rho_{\mathcal{R}}/(\sigma_k(A_\theta) - \rho)$: DDO's confusion loss $\mathcal{L}_{\text{confusion}}$ enlarges $\sigma_k(A_\theta)$ while the orthogonality constraint $\mathbf{u}_i \perp \hat{\mathbf{r}}_\ell$ keeps $\rho_{\mathcal{R}}$ small. The quantity $r_{\text{eff}}(\varepsilon) \triangleq \#\{j : \sigma_j(A_\theta) > \rho + \rho_{\mathcal{R}}/\varepsilon\}$ is the *estimated decoy phase budget*. When all neurons read the same direction $\hat{\mathbf{r}}_\ell$, A_θ is effectively rank 1. To raise the effective rank, DDO (rank k) splits the n edited neurons into k groups; each group g replaces $\hat{\mathbf{r}}_\ell$ with a randomly perturbed direction $\mathbf{q}_{\ell,g} = \text{normalize}(\hat{\mathbf{r}}_\ell + \gamma \mathbf{z}_{\ell,g})$ ($\mathbf{z}_{\ell,g} \perp \hat{\mathbf{r}}_\ell$ random, $\gamma > 0$ controls diversity), setting $\mathbf{w}_{\text{up},i} \leftarrow \mathbf{q}_{\ell,g(i)}$ and $\mathbf{w}_{\text{gate},i} \leftarrow \beta_g \mathbf{q}_{\ell,g(i)}$ for neurons i in group g.

Corollary 3 (Flat decoy spectra are optimal under fixed energy budget)**.** *If $\|A_\theta\|_F \leq B$ (decoy energy budget B), the largest achievable $\sigma_k(A_\theta)$ is $B/\sqrt{k}$, attained when $\sigma_1 = \cdots = \sigma_k = B/\sqrt{k}$.*

Concentrating decoy energy in one direction maximizes rank-1 protection but leaves higher ranks undefended; spreading it flat defends more ranks but each direction is weaker—this is the tradeoff that motivates DDO (rank k)'s grouping into k perturbed directions. Theorem 2 applies to fixed contrastive rank-k attacks; adaptive multi-phase RFA changes the contrast distribution after each phase, so r_{eff} serves as a diagnostic predictor rather than a formal guarantee for iterative re-estimation. These results scope to contrastive estimator attacks; weight-level attacks (Heretic) and a Stackelberg formalization are in Appendix E.2. Proofs are in Appendix E.1.

Orthogonal debiasing (utility repair). DDO can introduce mild over-refusal [27] on some models. We repair this with **orthogonal debiasing**: projecting out an over-refusal direction $\hat{\mathbf{v}}$ estimated from benign prompts the model incorrectly refuses [28, 29]. For matrices where the residual stream is the *input* dimension ($\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{up}}$): $\mathbf{W}' = \mathbf{W} - (\mathbf{W}\hat{\mathbf{v}})\hat{\mathbf{v}}^\top$. For matrices where the residual stream is the *output* dimension ($\mathbf{W}_o, \mathbf{W}_{\text{down}}$): $\mathbf{W}' = \mathbf{W} - \hat{\mathbf{v}}(\hat{\mathbf{v}}^\top \mathbf{W})$. We apply this to $\mathbf{W}_{\text{emb}}$, $\mathbf{W}_o$, and $\mathbf{W}_{\text{down}}$.

4 Experiments and Results

4.1 Experimental Setup

DDO edits are applied to model weights before deployment; the defended checkpoint incurs no architectural or runtime overhead. We compare against trained baselines under the same evaluation protocol. For all defenses, we report (i) utility and benign compliance without attack, and (ii) robustness under our attack ladder (standard RFA, adaptive multi-phase RFA, and Heretic; Section 2). Full settings, hyperparameters, and per-model configs are in Appendix F.2.

Table 2: Main defense comparison on Llama-3-8B-Instruct. ASR (%, 3-judge avg). MT-B = MT-Bench; XST = XSTest; DR = DirectRequest. RFA: 3p/rs = three-point/residual-stream; J/H = JailbreakBench/HarmBench. H-200 = Heretic at 200 trials. **best** / good / poor / ours. DDO matches trained baselines on standard RFA (1.8% avg) at 30–450× lower cost without finetuning.

		Utility ↑			Direct	RFA ASR ↓					Heretic ↓
Defense	Type	MMLU	MT-B	XST	DR↓	3p-J	rs-J	3p-H	rs-H	Avg	H-200
Base	–	**68.1**	**7.89**	94.8	3.8	86.0	84.0	87.2	83.2	85.1	88.7
Circuit Breakers	Trained	67.6	7.73	95.6	11.3	**0.0**	1.7	**0.0**	2.7	1.1	24.0
LAT	Trained	67.8	7.52	20.8	**0.0**	10.7	4.7	8.0	3.8	6.8	82.3
ReFAT	Trained	67.2	7.37	60.4	**0.0**	3.7	3.3	5.2	5.5	4.4	85.3
RepBend	Trained	64.4	7.69	98.4	**0.0**	6.0	4.7	4.0	4.4	4.8	**1.3**
Triplet	Trained	67.7	7.78	98.0	**0.0**	14.0	17.3	20.1	20.8	18.1	0.0
Triplet-Adv	Trained	65.6	7.47	96.8	25.6	8.7	3.7	12.2	6.3	7.7	30.0
DDO	Ours	66.8	7.38	91.6	0.4	2.3	0.3	4.4	**0.0**	1.8	18.0
DDO + debiasing	Ours	66.8	7.56	**99.2**	0.4	0.7	**0.0**	3.4	**0.0**	1.0	22.3

Models and baselines. Our primary evaluation uses Llama-3-8B-Instruct with the full attack ladder and baseline comparison. On Llama-3, we compare against six trained defenses using publicly released checkpoints: Circuit Breakers [9], LAT [10], ReFAT [11], RepBend [12], Triplet, and Triplet-Adv [13]. For cross-model generalization, we evaluate DDO on five additional instruction-tuned model families and compare against four trained baselines (Circuit Breakers, RepBend, LAT, ReFAT) with small hyperparameter sweeps (3–4 configs per defense per model; Appendix F.8). Seven further models are evaluated with DDO only (Appendix F.15).

Evaluation protocol. All generations use deterministic greedy decoding on a single A100 GPU. Hyperparameter tuning and DIM estimation use a held-out validation split (128 harmful + 128 safe probes); all reported metrics are on held-out test sets. We report knowledge accuracy (MMLU [36], 5-shot), conversational quality (MT-Bench [37]), over-refusal on prompts that superficially resemble harmful ones (XSTest [38], 250 such prompts, GPT-4o judge), and robustness under DirectRequest (harmful prompts sent directly), four standard-RFA variants (Section 2), adaptive multi-phase RFA (up to 8 phases), Heretic (200 trials), and prompt-level jailbreaks (HumanJailbreaks, GCG [21], PAIR [22], AutoDAN [23]). ASR is averaged over three judges: HarmBench classifier [5], LlamaGuard-2 [39], and StrongREJECT [40]. Configurations that fail a coherence check (non-empty, non-degenerate, minimum-length outputs on 10 benign prompts) are marked "×" and excluded. Full details on decoding, DIM estimation, splits, and benchmark settings are in Appendix F.7.

4.2 Results

Standard RFA and Heretic on Llama-3. Table 2 compares DDO against six trained baselines on Llama-3-8B-Instruct. **DDO + debiasing** adds orthogonal debiasing (Section 3) to repair over-refusal. On standard RFA, DDO achieves 1.8% mean ASR, comparable to the strongest trained baselines while requiring no finetuning and ∼2 minutes per checkpoint. Several trained defenses reduce ASR partly by over-refusing (LAT: 20.8% XSTest; ReFAT: 60.4%), while DDO preserves 91.6%. Adding orthogonal debiasing restores XSTest to 99.2% and improves standard RFA further (mean ASR 1.0%). Under Heretic [14] (weight-level attack), DDO achieves 18% H-200 ASR; a gap to trained RepBend (1.3%) remains, consistent with trained defenses distributing refusal more diffusely across parameters.

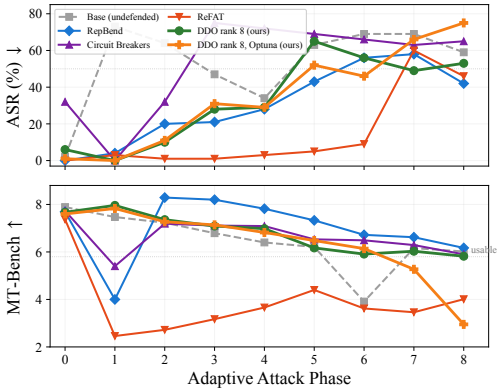

Figure 2: Adaptive RFA on Llama-3-8B (8 phases). **Top**: attack success rate (↓). **Bottom**: model quality (↑). All defenses see rising ASR under sustained re-estimation: DDO (green) remains coherent (MT-Bench ≥5.82) but ultimately leaks, while ReFAT (red) attains low ASR partly by collapsing quality to 2.46 at Phase 1.

Adaptive multi-phase RFA. Under adaptive multi-phase RFA (phases $k=1,\ldots,8$), all evaluated defenses degrade as the attacker re-estimates.

Worst-case ASR across phases is comparable across the strongest defenses: RepBend [12] 58%, ReFAT [11] 60%, DDO (rank 8) 65%, with Circuit Breakers [9] weaker at 75%. We track ASR jointly with utility (MT-Bench [37]; Figure 2) because ASR alone can be misleading—a defense can lower ASR by degrading generation rather than refusing. At Phase 8, DDO (rank 8) preserves MT-Bench at 5.82, comparable to RepBend (6.17) and Circuit Breakers (5.90), and substantially better than ReFAT, which collapses to 4.01 (Appendix Table 17). DDO therefore lands in the same operating regime as the strongest gradient-trained defenses on both worst-case ASR and utility under sustained adaptive attack, despite requiring no finetuning.

Cross-model generalization and baseline comparison. To test generalization, we apply DDO to five additional model families (Yi-1.5, Qwen3, Gemma-2, Mistral, and GLM-4) and compare against four trained baselines per model (Table 3; Appendix F.8).

Table 3: Cross-model baseline comparison (ASR %, 3-judge avg). DR = DirectRequest; RFA = mean of 4 variants (three-point/residual-stream × JailbreakBench/HarmBench). Cost: A100 wall-clock time per run (single config). **best** / good / poor / ours. DDO achieves the lowest RFA among utility-preserving defenses on 4/5 models at 30–450× lower cost.

Model	Defense	MMLU↑	MT-B↑	XST↑	DR↓	RFA↓	Cost
Yi-1.5-9B	Base	71.3	7.92	**98.0**	27.0	84.9	–
Yi-1.5-9B	Circuit Breakers	**71.4**	7.93	93.6	50.3	**0.0**	∼1h
Yi-1.5-9B	RepBend	71.1	**8.00**	94.0	44.3	**0.0**	∼5h
Yi-1.5-9B	LAT	71.3	1.00	0.4	**0.0**	**0.0**	∼15h
Yi-1.5-9B	ReFAT	70.4	6.60	90.0	0.3	**0.0**	∼8h
Yi-1.5-9B	**Ours (DDO)**	70.6	7.78	93.6	5.0	5.0	∼2m
Qwen3-8B	Base	76.6	6.78	94.8	8.5	90.7	–
Qwen3-8B	Circuit Breakers	76.6	6.85	98.0	10.7	77.6	∼1h
Qwen3-8B	RepBend (D)	76.7	6.78	96.8	0.3	24.8	∼5h
Qwen3-8B	LAT (C)	76.7	2.35	0.0	**0.0**	**0.0**	∼15h
Qwen3-8B	ReFAT (C)	**77.4**	6.12	97.2	8.5	8.4	∼8h
Qwen3-8B	**Ours (DDO)**	76.6	**7.15**	**100.0**	6.2	8.0	∼2m
Gemma-2-9B	Base	73.4	8.35	80.4	1.3	66.5	–
Gemma-2-9B	Circuit Breakers (B)	73.4	8.42	79.2	1.3	65.0	∼1h
Gemma-2-9B	RepBend (D)	72.5	8.17	**86.4**	**0.0**	5.5	∼5h
Gemma-2-9B	LAT	73.4	8.40	51.2	**0.0**	**0.0**	∼15h
Gemma-2-9B	ReFAT	73.4	7.32	77.6	**0.0**	7.4	∼8h
Gemma-2-9B	**Ours (DDO)**	**73.6**	**8.43**	83.2	2.1	**0.0**	∼2m
Mistral-7B	Base	60.1	7.51	94.0	40.6	86.3	–
Mistral-7B	Circuit Breakers (C)	59.4	7.47	**96.0**	**0.0**	35.2	∼1h
Mistral-7B	RepBend	60.0	7.51	94.0	18.9	81.0	∼5h
Mistral-7B	LAT (C)	57.5	**7.64**	63.6	**0.0**	4.2	∼15h
Mistral-7B	ReFAT (C)	**61.0**	6.58	**96.0**	41.5	57.6	∼8h
Mistral-7B	**Ours (DDO)**	60.1	7.51	92.0	5.7	**0.0**	∼2m
GLM-4-9B	Base	**70.3**	**7.78**	96.8	26.1	87.0	–
GLM-4-9B	Circuit Breakers (D)	70.1	7.73	96.0	8.5	6.5	∼1h
GLM-4-9B	RepBend	69.7	7.64	**98.4**	**0.0**	13.3	∼5h
GLM-4-9B	LAT (C)	70.1	7.67	84.4	5.7	59.2	∼15h
GLM-4-9B	ReFAT	69.2	2.84	95.6	18.2	52.9	∼8h
GLM-4-9B	**Ours (DDO)**	**70.3**	**7.78**	90.8	0.3	**2.0**	∼2m

Across all six models, DDO achieves $< 10\%$ mean standard-RFA ASR without base-model finetuning. On Qwen3 and GLM-4, DDO achieves the lowest RFA ASR among utility-preserving defenses. On Gemma-2, DDO matches LAT's 0% RFA while preserving higher benign compliance (83.2% vs. 51.2% XSTest). DDO also reduces DirectRequest compliance on models with weaker native refusal (Yi: $27\% \rightarrow 5\%$; Mistral: $41\% \rightarrow 6\%$; GLM-4: $26\% \rightarrow 0.3\%$). Beyond these six families, DDO transfers to seven additional models with per-model Optuna tuning (∼10 min each), achieving <10% ASR on all of them (Appendix F.15). DDO hardens a checkpoint in ∼2 minutes on a single A100 GPU, a 30–450× cost advantage over trained baselines (1–15 A100-hours each).

Compile-mode selection. For each model, we sweep both compile modes (replace and additive) with 15 Optuna trials per mode (Table 10; Appendix Table 13). Replace mode overwrites the neuron's original weights with the decoy, producing a stronger signal but removing the neuron's original computation; additive mode superposes the decoy on top, preserving the original behavior at the cost

of a weaker decoy. The preferred mode depends on how each model's refusal circuitry responds to neuron overwriting: replace is preferred on Yi, Llama-3, and GLM-4 (where the model compensates for the lost neuron), while additive is preferred on Gemma-2 (replace fails the coherence check), Qwen3, and Mistral (replace disrupts baseline refusal). See Appendix F.4 for a detailed analysis.

Prompt-level jailbreaks. Although DDO targets mechanistic refusal removal, we also evaluate it against prompt-level attacks that operate purely at the input level (Table 4). DDO + debiasing achieves 0.4% GCG [21] ASR, matching the best trained defenses, and 1.2% on HumanJailbreaks. A plausible explanation is that optimization-based prompt attacks like GCG suppress the same low-dimensional refusal feature targeted by RFA [11]; DDO's decoy neurons partially re-inject refusal-correlated signal, making it harder for gradient-based prompt optimization to fully suppress refusal. However, semantic attacks like PAIR [22] (36.3% ASR) circumvent refusal through meaning rather than activation geometry, and DDO provides no advantage here. Defending against semantic jailbreaks requires complementary input-level defenses such as input classifiers.

Table 4: Prompt-level jailbreak ASR (%↓, 3-judge avg, HarmBench). HJB = HumanJailbreaks; ADAN = AutoDAN. **best** / good / poor / ours. DDO + debiasing matches the best trained defenses on GCG (0.4%); PAIR remains high (36.3%).

Defense	Type	HJB↓	GCG↓	PAIR↓	ADAN↓
Base	–	2.9	27.7	55.1	0.4
CB	Trained	8.2	3.8	32.1	5.7
LAT	Trained	**0.0**	8.0	34.6	**0.0**
ReFAT	Trained	**0.0**	30.6	**30.0**	**0.0**
RepBend	Trained	0.1	0.6	38.8	0.2
Triplet	Trained	**0.0**	**0.4**	39.8	**0.0**
Tri-Adv	Trained	20.9	22.2	43.6	20.8
DDO	Ours	1.2	1.7	37.5	**0.0**
+debias	Ours	1.2	**0.4**	36.3	**0.0**

4.3 Ablations

Table 5 isolates each DDO component on Llama-3-8B-Instruct by removing one element at a time from the full configuration. All ablations preserve MMLU within 3 points of the base model.

Gradient optimization is essential. Random orthogonal decoys (no optimization) achieve 72.3% ASR—only a modest reduction from the undefended base (85.1%). In contrast, gradient optimization under the full 4-part objective reduces ASR to 1.8%, confirming that orthogonality alone is insufficient: decoys must be optimized to dominate the harmful–safe contrast that the attacker re-estimates.

Loss components. The *confusion loss* $\mathcal{L}_{\text{confusion}}$ and *refusal-score loss* $\mathcal{L}_{\text{refusal-score}}$ are the most critical terms: removing either increases ASR to 43.2% and 41.8% respectively. The confusion loss directly optimizes decoy directions to mislead the DIM estimator, while the refusal-score loss prevents first-token collapse ("Sure ..." followed by refusal). The retain KL term stabilizes optimization (ASR 22.4% without it) but is less critical on its own.

Neuron selection and compile mode. Low-norm neuron selection outperforms random neuron choice (ASR 1.8% vs. 15.3%), supporting our design choice to edit low-impact units to reduce utility disruption while maintaining decoy strength. On Llama-3, replace mode outperforms additive (ASR 1.8% vs. 16.5%), but this preference is model-specific (Appendix F.4). Finally, layer placement upstream of the causal refusal zone is critical on models with localized refusal (Appendix F.5).

Table 5: DDO ablation on Llama-3-8B-Instruct. RFA ASR = mean of 4 variants, 3-judge avg (%). **best** / good / poor / ours. All components contribute; confusion and refusal-score losses are most critical.

Ablation	RFA↓	MMLU↑	XST↑
Base (no defense)	85.1	**68.1**	94.8
Random orth. decoys	72.3	67.9	**99.6**
w/o confusion loss	43.2	66.5	82.0
w/o refusal-score	41.8	66.7	90.0
w/o retain KL	22.4	65.8	86.4
Additive (vs. replace)	16.5	66.8	82.8
Random neurons	15.3	65.2	79.2
DDO full	**1.8**	66.8	91.6

Failed approaches. We evaluated 14 additional post-hoc mechanisms beyond DDO (Appendix Table 19). Linear defenses are brittle under re-estimation: Decoy Shear Transform (92% ASR), Refusal Direction Rotation (68%), and Representation Rerouting (86%) are all defeated once the attacker recomputes DIM on the defended checkpoint. Aggressive edits (e.g., LM-Head Row Scaling) can drive RFA ASR near zero but at catastrophic utility cost (XSTest 8%, MT-Bench 1.72). Finally, spreading decoy energy across more directions involves a tradeoff predicted by Corollary 3: diversified readers weaken rank-1 protection (ASR rises from 10% to 27% as the number of decoy groups k increases from 1 to 4) but strengthen higher-rank robustness (rank-16 ASR drops from 5% to 2%). The optimal k depends on the expected attacker rank budget (Appendix E.3).

Cross-architecture mechanism comparison. We evaluated 8 post-hoc mechanisms across all 7 models (10–30 Optuna trials each; Appendix F.12). DDO is the only single mechanism that achieves $< 10\%$ ASR reliably across all architectures. Other mechanisms are individually weak (e.g., gate_boost: 56–81% ASR) or break generation on most models (diverse Q-head routing,

rotation, V-projection, head amplification, and KV strengthening all fail the coherence check on Yi, Gemma-2, and GLM-4). Even random orthogonal decoys without gradient optimization fail on 5 of 7 models (18–53% ASR). This supports the core DDO design: nonlinear SwiGLU gating plus gradient optimization (and per-model compile-mode selection) are jointly necessary for cross-architecture robustness. Full mechanism definitions in Appendix D.2).

4.4 Discussion

The role of nonlinearity. Adaptive multi-phase RFA repeatedly fits a *linear* DIM direction on the defended checkpoint. Any defense equivalent to an input-independent linear map (e.g., a shear or rotation) remains removable after re-estimation. DDO instead introduces a prompt-dependent nonlinearity via SwiGLU gating: writing $a \triangleq \langle \hat{\mathbf{r}}_\ell, \tilde{\mathbf{x}}_\ell \rangle$, the decoy amplitude scales as $\mathrm{silu}(\beta a) \cdot a$ and behaves like βa^2 for large positive a. This creates harmful-selective mean shifts that corrupt the contrastive estimator but cannot be undone by a fixed linear projection.

Resistance to DDO undo attacks. A defense-aware attacker may try to *undo* DDO by identifying edited neurons. Let $\hat{\mathbf{r}}_{\mathrm{clean}}$ denote the true refusal direction and $\hat{\mathbf{r}}_{\mathrm{atk}}$ the DIM direction the attacker estimates on the defended checkpoint. The attacker scores each MLP gate row by $s_i \triangleq |\langle \mathbf{w}_{\mathrm{gate}}^{(i)}, \hat{\mathbf{r}}_{\mathrm{atk}} \rangle|$ and zeros the top-scoring rows. This would work if $\hat{\mathbf{r}}_{\mathrm{atk}} \approx \hat{\mathbf{r}}_{\mathrm{clean}}$, since DDO writes $\hat{\mathbf{r}}_{\mathrm{clean}}$-aligned triggers into gate rows. Instead, estimator corruption rotates $\hat{\mathbf{r}}_{\mathrm{atk}}$ away from $\hat{\mathbf{r}}_{\mathrm{clean}}$, so edited neurons are not salient under this score. Across Qwen3, Mistral-7B, and Yi, mid-zone target layers satisfy $\mathrm{cossim}(\hat{\mathbf{r}}_{\mathrm{clean}}, \hat{\mathbf{r}}_{\mathrm{atk}}) \approx 0$ and the modified rows' scores fall within the natural top-5 distribution. This undo heuristic therefore offers no advantage over standard RFA; post-attack ASR under three undo heuristics confirms this (Appendix F.16). DDO's robustness is also not sample-limited: ASR remains flat at 6–8% as the attacker's probe budget increases from 32 to 1024 (Appendix F.18). Under rank-k SVD attacks, DDO with a single reader is effective at rank 1 (4% ASR) but degrades at higher ranks (39% at k=16; Appendix F.17). Diversifying readers across K groups extends higher-rank protection at the cost of rank-1 strength, as predicted by the spectral toll (Appendix E.3).

Comparison with trained defenses. On standard RFA, DDO matches the strongest trained baselines across five model families (Table 3) at 30–450× lower cost. Under Heretic on Llama-3, DDO alone (18%) remains weaker than RepBend [12] (1.3%). A plausible explanation is *entanglement*: gradient-based training can distribute refusal across many parameters jointly, while closed-form edits are easier to localize and therefore easier to target with weight-level attacks.

5 Conclusion and Limitations

What DDO changes about the defense landscape. Existing defenses against refusal ablation require gradient-based training per checkpoint, creating a bottleneck for the open-weight release cycle. DDO demonstrates that a mechanistic alternative—repurposing MLP neurons to corrupt the attacker's contrastive estimator—can match trained baselines on standard RFA across six model families at 30–450× lower cost. The key insight is that contrastive attacks are only as good as the contrast they estimate: by injecting a large, harmful-selective, refusal-orthogonal decoy signal, DDO forces the attacker to ablate decoys rather than genuine refusal.
DDO requires only a small generic probe set (128 harmful + 128 safe prompts), gradient optimization of the decoy parameters (base weights frozen, so memory overhead beyond inference is minimal), and a single GPU—no full finetuning infrastructure, no method-specific safety datasets, and few design decisions beyond compile mode and layer range. This makes it applicable by any party with access to the weights: model providers before release, downstream deployers, safety auditors, or automated agents. The feedback loop is fast (minutes per checkpoint) and defense strength is controllable via compile mode, and scalar gains (β, s), making DDO well-suited to agent-driven safety pipelines that iteratively harden and evaluate checkpoints.

Limitations and future work. DDO does not make refusal unremovable. Sustained adaptive re-estimation challenges all evaluated defenses; and DDO can introduce over-refusal on some models, though orthogonal debiasing and compile-mode selection mitigate this (Appendix A). Concrete directions for future work include: developing post-hoc mechanisms that induce deeper parameter–behavior entanglement to close the gap with trained defenses under weight-level attacks; and studying DDO with composed with inference-time guardrails (e.g., Llama Guard) for defense-in-depth.

References

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[3] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[4] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229, 2022.

[5] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[6] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[7] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

[8] Bahrad A. Sokhansanj. Uncensored AI in the wild: Tracking publicly available and locally deployable LLMs. *Future Internet*, 17(10):477, 2025.

[9] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37: 83345–83373, 2024.

[10] Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*, 2024.

[11] Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust LLM safeguarding via refusal feature adversarial training. *arXiv preprint arXiv:2409.20089*, 2024.

[12] Ashkan Yousefpour, Taeheon Kim, Ryan Sungmo Kwon, Seungbeen Lee, Wonje Jeung, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, and Jonghyun Choi. Representation bending for large language model safety. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24073–24098, 2025.

[13] Samuel Simko, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. Improving large language model safety with contrastive representation learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28154–28182, 2025.

[14] Philipp Emanuel Weidmann. Heretic: Fully automatic LLM abliteration. `https://github.com/p-e-w/heretic`, 2025. Open-source tool, AGPL-3.0.

[15] Luke Emberson. Open-weight models lag state-of-the-art by around 3 months on average, 2025. URL `https://epoch.ai/data-insights/open-weights-vs-closed-weights-models/`. Accessed: 2026-05-01.

[16] Jean-Stanislas Denain. Models with downloadable weights currently lag behind the top-performing models, 2025. URL `https://epoch.ai/data-insights/open-vs-closed-model-performance`. Accessed: 2026-05-01.

[17] Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

[18] Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There is more to refusal in large language models than a single direction. *arXiv preprint arXiv:2602.02132*, 2026.

[19] Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. *arXiv preprint arXiv:2502.09674*, 2025.

[20] Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025.

[21] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[22] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025.

[23] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[24] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

[25] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[26] Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C. Wallace, and David Bau. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=MxbEiFRf39`.

[27] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 11515–11542. PMLR, 2025.

[28] Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. *arXiv preprint arXiv:2410.03415*, 2024.

[29] Mahavir Dabas, Si Chen, Charles Fleming, Ming Jin, and Ruoxi Jia. Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 11846–11861. PMLR, 2025.

[30] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[31] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024.

[32] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Wang, Bowen Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[33] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Cassirer, Siamak Barber, Armand Joulin, Marc'Aurelio Ranzato, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[34] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[35] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Roessler, et al. GLM-4: Open bilingual pre-trained model. *arXiv preprint arXiv:2406.12793*, 2024.

[36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[37] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[38] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, 2024.

[39] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[40] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

[41] Maxime Labonne. Uncensor any LLM with abliteration. `https://huggingface.co/blog/mlabonne/abliteration`, 2024. Hugging Face blog.

[42] Stephen Casper, Kyle O'Brien, Shayne Longpre, Elizabeth Seger, et al. Open technical problems in open-weight AI model risk management. TMLR / OpenReview, 2025. URL `https://openreview.net/forum?id=8QyGLnFkzc`.

[43] National Institute of Standards and Technology. Managing misuse risk for dual-use foundation models. Technical Report NIST AI 800-1, Second Public Draft, National Institute of Standards and Technology, 2025. URL `https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf`.

[44] Chen Bo Calvin Zhang, Christina Q. Knight, Nicholas Kruus, et al. LLM novice uplift on dual-use, in silico biology tasks. *arXiv preprint arXiv:2602.23329*, 2026.

[45] Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz S. Hanke, et al. Dual-use capabilities of concern of biological AI models. *PLOS Computational Biology*, 21(5):e1012975, 2025. doi: 10.1371/journal.pcbi.1012975.

[46] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28525–28550. PMLR, 2024.

[47] Caoilte Ó Ciardha, John Buckley, and Rebecca Portnoff. AI-generated child sexual abuse material: What's the harm? *AI & Society*, 2026. doi: 10.1007/s00146-026-02932-y.

[48] Internet Watch Foundation. AI-generated images. Technical report, Internet Watch Foundation, 2025. URL `https://www.iwf.org.uk/annual-data-insights-report-2025/image-insights/ai-generated-images/`.

[49] Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. *arXiv preprint arXiv:2406.14144*, 2024.

[50] Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to LLM security. *arXiv preprint arXiv:2408.17003*, 2024.

[51] Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488, 2024.

[52] Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Roy Ka-Wei Lee, Erik Cambria, and Ranjan Satapathy. Understanding refusal in language models with sparse autoencoders. 2025.

[53] Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*, 2024.

[54] Samyak Jain, Ekdeep S Lubana, Kemal Oksuz, Tom Joy, Philip H Torr, Amartya Sanyal, and Puneet K Dokania. What makes and breaks safety fine-tuning? a mechanistic study. *Advances in Neural Information Processing Systems*, 37:93406–93478, 2024.

[55] Abhay Sheshadri, Aidan Ewart, Phillip Huang Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. *arXiv preprint arXiv:2407.15549*, 2024.

[56] Harethah Abu Shairah, Hasan Abed Al Kader Hammoud, Bernard Ghanem, and George Turkiyyah. An embarrassingly simple defense against LLM abliteration attacks. *arXiv preprint arXiv:2505.19056*, 2025.

[57] Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. *arXiv preprint arXiv:2505.17306*, 2025.

[58] Richard J Young. Comparative analysis of LLM abliteration methods: A cross-architecture evaluation. *arXiv preprint arXiv:2512.13655*, 2025.

[59] Paul Youssef, Zhixue Zhao, Daniel Braun, Jörg Schlötterer, and Christin Seifert. Position: Editing large language models poses serious safety risks. *arXiv preprint arXiv:2502.02958*, 2025.

[60] Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[61] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

A Limitations and Future Work

We provide a broad evaluation within the attack ladder studied here (three attacker tiers, six primary model families with trained baselines, seven additional models with DDO only, and six trained baselines on Llama-3), but several open problems remain.

Adaptive attackers. Sustained adaptive re-estimation remains an open challenge for the field: *no* evaluated defense—trained or post-hoc—achieves low ASR beyond 8 phases. DDO (rank 8) reaches 65% worst-case ASR (LlamaGuard-2; Table 16), comparable to trained RepBend's, indicating that the bottleneck is not training versus post-hoc editing but the fundamental difficulty of defending against iterative re-estimation. Future work should evaluate stronger adaptive attackers (e.g., longer phase budgets, joint multi-direction projections, and alternative estimators beyond DIM).

Weight-level attacks. DDO achieves 18% Heretic ASR, outperforming most trained baselines (Circuit Breakers, LAT, ReFAT) that are fully compromised under the same attack (Table 2). RepBend [12] achieves lower ASR through gradient-based parameter–behavior entanglement that is difficult to replicate with post-hoc edits alone. Closing this gap—e.g., via post-hoc edits that induce deeper entanglement surviving weight-level optimization—is a promising direction for future work.

Over-refusal on some models. Over-refusal is a well-known trade-off across safety-hardening methods, and DDO is no exception: some models show reduced benign compliance. Notably, several trained baselines exhibit comparable or worse degradation (e.g., LAT and ReFAT on Llama-3; Table 2), suggesting this is a general challenge rather than a post-hoc limitation. DDO already provides two effective mitigations: (i) orthogonal debiasing repairs benign compliance without sacrificing RFA robustness (Section 3), and (ii) switching to additive mode reduces over-refusal on some architectures. Automatically selecting the best configuration per model remains an open direction.

Compile mode selection. Selecting between replace and additive compile modes currently requires running both and comparing ASR/XSTest. This search is already inexpensive relative to training (minutes versus hours), but a predictive heuristic based on refusal redundancy could further streamline deployment.

A.1 Broader Impacts

The threat is operational, not hypothetical. Training-free refusal removal has moved well beyond academic proof-of-concept. The mechanistic result that refusal is mediated by a low-dimensional residual-stream direction [7] was rapidly operationalized into public tooling: community tutorials describe one-command uncensoring without retraining [41], and automated tools such as Heretic apply hyperparameter-optimized weight edits on a consumer GPU with no ML expertise required [14]. A recent large-scale audit of public model repositories found over 8,600 safety-modified open-weight checkpoints across 1,300 namespaces, with more than half packaged in GGUF format optimized for local consumer-hardware deployment and accumulating tens of millions of tracked downloads [8]. Refusal removal has therefore become a distribution and hosting problem: users can either run abliteration scripts themselves, or simply download an already-modified checkpoint.

DDO as a platform-side hardening primitive. This distribution surface is exactly where DDO intervenes. Once aligned weights are publicly downloadable, there is no practical way to guarantee that all copies receive safety updates or can be rolled back [42]. Per-checkpoint safety retraining cannot address this at scale: it is too slow and too expensive to apply to every community finetune, merge, or re-quantized derivative hosted on a platform. DDO—running in two minutes on a single GPU with no training data—can be applied by any party in the distribution chain before serving or redistributing a checkpoint. This is consistent with emerging policy guidance: NIST explicitly assigns monitoring and mitigation responsibilities to model-hosting platforms and other distribution channels [43]. By raising the cost of the most accessible refusal-removal workflows, DDO addresses a gap that neither upstream safety training nor downstream prompt filtering can close alone.

Harm domains where cost-raising matters most. The impact of increasing the attacker's cost is largest in domains where compliant local assistance lowers barriers to serious misuse. For CBRN and biosecurity, LLM access has been shown to substantially accelerate novices on dual-use biology tasks:

a controlled uplift study found that participants with LLM assistance were $4.16\times$ more accurate than internet-only controls, with 89.6% reporting little difficulty obtaining dual-use information despite safeguards [44]. Pre-deployment evaluation of high-consequence biological capabilities is now considered a baseline safety requirement [45], and WMDP provides a systematic benchmark for hazardous biosecurity and chemical knowledge in LLMs [46]. For child safety, robust refusal reduces the availability of uncensored local assistants that can be used to facilitate grooming, coercion, or integration into abuse pipelines. Academic analysis identifies AI-generated CSAM as enabling revictimization, normalization, and lowered barriers to offending [47], while operational reporting documents sharply rising volumes and calls for safety-by-design action by platform providers [48]. In both domains, DDO does not provide complete protection—determined adversaries with sufficient resources can still bypass it—but it raises the cost of the most accessible and widespread attack workflows.

Potential negative impacts and dual-use concerns. Three dual-use risks merit discussion. First, our adaptive multi-phase RFA attack is a new, stronger variant of existing RFA; however, it is a natural extension of publicly available methods, and we believe the defensive contribution outweighs this incremental offensive capability. Second, our mechanistic analysis of why defenses fail (nonlinearity requirements, spectral toll, undo attack camouflage) could inform adversaries seeking to circumvent safety measures; we mitigate this by ensuring all analysis serves the defensive design and by not releasing attack-only tools. Third, DDO should be understood as a cost-raising hardening layer—not a complete safety guarantee—and we report its residual attack surfaces transparently in Section A.

Responsible disclosure. We do not release new uncensored model checkpoints, finetuning recipes for removing safety guardrails, or standalone attack tools. DDO code is available as a defensive hardening tool. All evaluation uses established public benchmarks (JailbreakBench, HarmBench, AdvBench) and publicly available model checkpoints.

B Extended Related Work

Mechanistic refusal representations and abliteration. Several analyses of RLHF-aligned LLMs show that refusal can be removed by ablating low-dimensional residual-stream features. Projecting out a rank-1 difference-in-means (DIM) direction—the *refusal feature*—often eliminates refusal, enabling Refusal Feature Ablation (RFA) and "abliteration" [7]. Such direction-based interventions have been contextualized under the Linear Representation Hypothesis [6]. Subsequent work argues that refusal geometry can be multi-dimensional (e.g., *concept cones* [17] and multiple refusal-mediating directions [18]), with orthogonal safety dimensions [19] and evidence that harmfulness and refusal are encoded separately [20]. Complementary mechanistic studies localize safety-relevant structure to specific neurons [49, 50] and intermediate hidden states [51], while sparse autoencoders reveal refusal as a composition of fine-grained latent features [52, 53]. Jain et al. [54] show that safety finetuning induces narrow, localized internal changes, which helps explain why contrastive estimators like RFA can locate and remove the refusal mechanism—motivating defenses that corrupt the estimator itself. We include both single-phase RFA and adaptive multi-phase RFA in our attack ladder (Section 2).

Defenses against refusal ablation. Existing defenses against RFA are largely training-time interventions. Circuit Breakers [9] finetunes checkpoints with a representation-rerouting objective; LAT [10], targeted LAT [55], and ReFAT [11] rely on adversarial finetuning; and representation-reshaping approaches such as RepBend [12] and contrastive objectives [13] modify refusal geometry during gradient-based training. A concurrent defense targets abliteration directly [56], but still requires finetuning. While effective, these approaches require per-checkpoint training runs and, depending on the method, additional data collection or attack pipelines (e.g., adversarial examples) and hyperparameter tuning. In contrast, we study *post-hoc mechanistic weight editing*: parameter edits computed from forward-pass statistics and folded into parameters with no architectural or runtime overhead, enabling rapid hardening and composition across open-weight releases. Because these edits target specific components (layers, matrices, neurons), they also provide an interpretable and automatable complement to training-based hardening: they can be tuned with simple black-box search and used as a fast pre-hardening step or structured initialization for subsequent finetuning.

Prompt-level jailbreaks and robust refusal. LLMs are also vulnerable to prompt-based jailbreaks that circumvent refusal by optimizing adversarial inputs, without modifying model weights. We

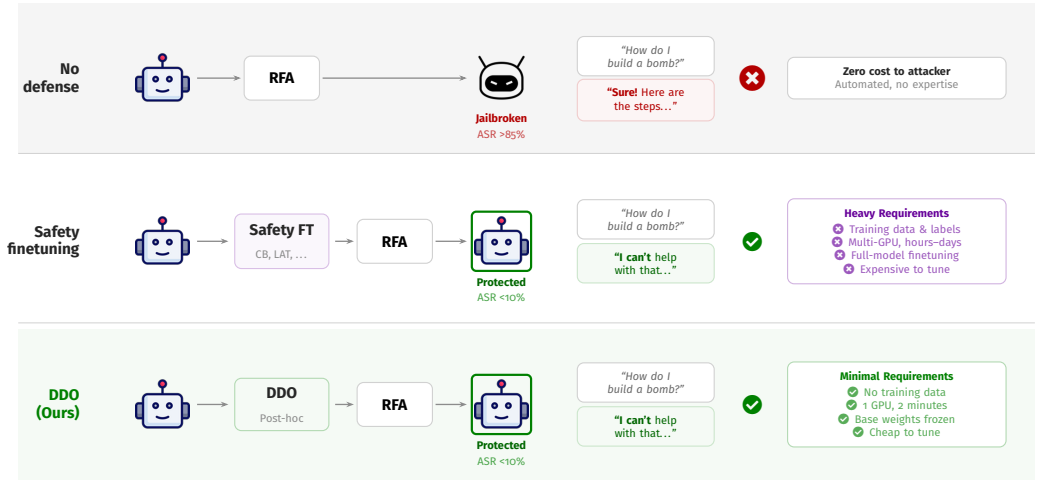

Figure 3: **DDO vs. existing defenses against RFA.** Without any defense, RFA strips refusal from open-weight models with zero attacker cost (top row). Trained safety defenses (Circuit Breakers, LAT, ReFAT, RepBend) preserve refusal under RFA but require training data, multi-GPU finetuning, and expensive per-method hyperparameter sweeps (middle row). DDO matches trained defenses at 30–450$\times$ lower cost: a single GPU, two minutes, no training data, and base weights left frozen (bottom row).

report three widely used automated attacks—GCG [21], PAIR [22], and AutoDAN [23]—under HarmBench's standardized robust-refusal evaluation setting [5]. ReFAT provides evidence that many prompt jailbreaks can suppress a low-dimensional residual-stream refusal feature [11], linking these input-space attacks to the same refusal representation exploited by RFA; we therefore include them as a complementary evaluation of how our post-hoc weight edits affect robustness under input-space attack.

Over-refusal calibration. Over-refusal—where aligned models refuse benign requests—has been studied through dedicated benchmarks [27] and addressed via representation-level interventions. A single learned direction can calibrate false refusal in a training-free manner [28], while targeted representation finetuning can shift refusal boundaries with minimal utility loss [29]. Our orthogonal debiasing mechanism (Section 3) is complementary: it projects out an over-refusal direction from weight matrices after DDO, restoring benign compliance while preserving robustness.

Real-world prevalence of abliteration. The refusal direction exploited by RFA transfers across languages [57], broadening the scope of the vulnerability beyond English-only settings. A large-scale audit found over 8,600 uncensored model repositories on HuggingFace, with abliterated variants achieving $\sim$74–80% compliance on unsafe prompts [8]. The growing ecosystem of automated abliteration tools across architectures has been surveyed [58], and cheap post-hoc weight editing has been argued to pose serious safety risks for open-weight releases [59]. These considerations motivate our focus on post-hoc hardening primitives that can be applied rapidly across checkpoints.

C Notation and Background

We use a standard decoder-only transformer architecture (e.g., Llama-3 [30]). For a fixed token position (we omit token indices), let $\mathbf{h}_\ell \in \mathbb{R}^d$ denote the residual-stream state at the input of layer ℓ (hidden dimension d). Each layer is *pre-norm*: RMSNorm is applied before attention and before the MLP, and each sublayer update is added back to the residual stream. We write RMSNorm with gain $\boldsymbol{\gamma}_\ell \in \mathbb{R}^d$ as $\mathrm{RMSNorm}(\mathbf{x}) \triangleq \mathbf{x}/\mathrm{RMS}(\mathbf{x}) \odot \boldsymbol{\gamma}_\ell$, where $\mathrm{RMS}(\mathbf{x}) \triangleq \sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2 + \varepsilon}$.

Attention. Let $\mathbf{x}_\ell \triangleq \mathrm{RMSNorm}(\mathbf{h}_\ell)$. Multi-head attention computes queries, keys, and values via learned projections:

$$\mathbf{Q}_\ell = \mathbf{W}_q \mathbf{x}_\ell, \quad \mathbf{K}_\ell = \mathbf{W}_k \mathbf{x}_\ell, \quad \mathbf{V}_\ell = \mathbf{W}_v \mathbf{x}_\ell, \tag{5}$$

Multi-head attention applies rotary positional embeddings (RoPE) to $\mathbf{Q}_\ell, \mathbf{K}_\ell$ and produces an output $\mathbf{a}_\ell$ projected back to the residual stream:

$$\mathbf{a}_\ell \triangleq \mathbf{W}_o \operatorname{Attn}(\mathbf{Q}_\ell, \mathbf{K}_\ell, \mathbf{V}_\ell), \qquad \mathbf{h}_\ell^{\mathrm{attn}} \triangleq \mathbf{h}_\ell + \mathbf{a}_\ell. \tag{6}$$

MLP (SwiGLU). Let $\tilde{\mathbf{x}}_\ell \triangleq \operatorname{RMSNorm}(\mathbf{h}_\ell^{\mathrm{attn}})$. Many modern LLMs (including Llama-3) use a SwiGLU MLP [60] with gate/up projections $\mathbf{W}_{\mathrm{gate}}, \mathbf{W}_{\mathrm{up}} \in \mathbb{R}^{d_{\mathrm{inter}} \times d}$ and down projection $\mathbf{W}_{\mathrm{down}} \in \mathbb{R}^{d \times d_{\mathrm{inter}}}$:

$$\begin{aligned} \mathbf{m}_\ell &\triangleq \mathbf{W}_{\mathrm{down}}\big(\operatorname{silu}(\mathbf{W}_{\mathrm{gate}}\tilde{\mathbf{x}}_\ell) \odot (\mathbf{W}_{\mathrm{up}}\tilde{\mathbf{x}}_\ell)\big), \\ \mathbf{h}_{\ell+1} &\triangleq \mathbf{h}_\ell^{\mathrm{attn}} + \mathbf{m}_\ell, \end{aligned} \tag{7}$$

where d_{inter} is the MLP intermediate dimension and $\odot$ is elementwise multiplication. Some models (e.g., Gemma-2) use GeGLU, which replaces silu with gelu; our method applies to both variants.

Weights we edit. Our defenses perform *offline* edits to selected parameter matrices, including $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o, \mathbf{W}_{\mathrm{gate}}, \mathbf{W}_{\mathrm{up}}, \mathbf{W}_{\mathrm{down}}$, the embedding matrix $\mathbf{W}_{\mathrm{emb}}$, the output head $\mathbf{W}_{\mathrm{lm}}$, and (for normalization-based mechanisms) RMSNorm gains $\{\boldsymbol{\gamma}_\ell\}$. We fold all edits into these weights before deployment, so defended models run with no architectural or runtime overhead.

Hookpoints (for attacks). When defining RFA variants (Section 2), we use three hookpoints per layer: residual-stream input $\mathbf{h}_\ell$, attention update $\mathbf{a}_\ell$, and MLP update $\mathbf{m}_\ell$.

D Additional Method Details

D.1 First-Token Collapse

During DDO optimization, the cross-entropy refusal loss $\mathcal{L}_{\text{refuse}}$ anchors safety by training the model to produce refusal continuations on harmful prompts. However, this loss operates on the full sequence and can be satisfied by degenerate solutions: the model outputs a compliance-prefixed first token (e.g., "Sure") to reduce the decoy-confusion loss, then pivots mid-sentence to refusal (e.g., "Sure, I'd be happy to... actually, I cannot assist with that request"). This "first-token collapse" is problematic because (i) autoregressive generation commits to the first token, so a compliance prefix can cascade into full compliance under greedy or sampled decoding, and (ii) ASR judges that inspect only the first few tokens may classify such responses as compliant even though the model eventually refuses.

The refusal-score loss $\mathcal{L}_{\text{refusal-score}}$ addresses this by directly supervising first-token logits:

$$\mathcal{L}_{\text{refusal-score}} = \max\left(0,\ m - \left[\max_{t \in \mathcal{T}_{\text{ref}}} z_t - \max_{t \in \mathcal{T}_{\text{comp}}} z_t\right]\right), \tag{8}$$

where z_t is the logit for token t, $\mathcal{T}_{\text{ref}} = \{\texttt{I}, \texttt{Sorry}, \texttt{cannot}\}$ are refusal-prefixed tokens, $\mathcal{T}_{\text{comp}} = \{\texttt{Sure}, \texttt{Here}\}$ are compliance-prefixed tokens, and m is a margin (m=0 in all experiments). This hinge loss ensures that refusal tokens dominate compliance tokens at the first position. In our ablation (Table 5), removing $\mathcal{L}_{\text{refusal-score}}$ increases ASR from 1.8% to 41.8%, confirming that first-token anchoring is essential for DDO's effectiveness.

D.2 All Mechanism Definitions

Section 3 defines DDO (the core mechanism). Here we define the 14 additional mechanisms that appear in at least one reported configuration, ablation, or negative-result analysis. Throughout, let $\mathbf{x}_\ell \triangleq \operatorname{RMSNorm}(\mathbf{h}_\ell)$ and $\tilde{\mathbf{x}}_\ell \triangleq \operatorname{RMSNorm}(\mathbf{h}_\ell^{\mathrm{attn}})$.

Decoy Shear Transform (linear baseline). We apply a linear shear that preserves the refusal coordinate but injects orthogonal decoy components: $\mathbf{M}_\ell \triangleq \mathbf{I} + \sum_j c_j\, \mathbf{u}_{\ell,j}\hat{\mathbf{r}}_\ell^\top$ with $\mathbf{u}_{\ell,j} \perp \hat{\mathbf{r}}_\ell$, folded into $\mathbf{W}_o$ and $\mathbf{W}_{\mathrm{down}}$ by left-multiplication. When harmful prompts induce larger $\langle \hat{\mathbf{r}}_\ell, \mathbf{h} \rangle$ than safe prompts, this shear adds a harmful-selective shift in decoy directions and can rotate the initial DIM estimate away from the causal refusal feature. However, because the transformation is linear, an adaptive attacker can re-estimate DIM on the defended checkpoint and recover the new contrast; on Llama-3 this yields 92% ASR (3-point RFA, LlamaGuard-2).

Diverse Q-Head Routing. For each attention head h, we define a head-specific trigger $\hat{\mathbf{t}}_{\ell,h} = \cos(\theta)\hat{\mathbf{r}}_\ell + \sin(\theta)\hat{\mathbf{o}}_{\ell,h}$ with $\hat{\mathbf{o}}_{\ell,h} \perp \hat{\mathbf{r}}_\ell$, and apply the rank-1 update

$$\mathbf{W}_{q,\ell}^{(h)} \leftarrow \mathbf{W}_{q,\ell}^{(h)} + \alpha(\mathbf{W}_{q,\ell}^{(h)}\hat{\mathbf{t}}_{\ell,h})\hat{\mathbf{t}}_{\ell,h}^\top. \tag{9}$$

This makes different heads respond to slightly different rotations of refusal, aiming to distribute harmful–safe separation across head-specific subspaces. Since attention includes a softmax nonlinearity, such diversification could in principle make the overall behavior harder to summarize by a single residual-stream DIM direction; in practice the edit is brittle and often fails the coherence check on some architectures.

RoPE Dimension Amplification. We rescale selected rotary frequencies $\omega_{\ell,j} \leftarrow \rho\,\omega_{\ell,j}$ for $j \in \mathcal{J}$, $\rho > 1$, by editing the RoPE parameters used by $\mathbf{W}_q$ and $\mathbf{W}_k$. If refusal-related computation relies on consistent attention patterns across positions, increasing rotation rates in a subset of dimensions can induce more position-dependent variation, potentially smearing harmful–safe differences across tokens and weakening a simple DIM estimator. This mechanism is highly architecture-sensitive and often degrades generation.

Value-Projection Conditioning. We add a rank-1 component to the value projection, $\mathbf{W}_{v,\ell} \leftarrow \mathbf{W}_{v,\ell} + \alpha\,\hat{\mathbf{u}}_\ell\hat{\mathbf{r}}_\ell^\top$, where $\hat{\mathbf{u}}_\ell$ is a fixed random unit vector. This explicitly routes the refusal coordinate $\langle\hat{\mathbf{r}}_\ell, \mathbf{x}\rangle$ into the value stream, so that even if an attacker partially removes refusal by ablating a mis-estimated direction, any surviving refusal signal is amplified and broadcast through attention aggregation downstream. This edit is attention-sensitive and can break coherence on some models.

KV Rank-1 Strengthening. This uses the same update as Value-Projection Conditioning but with a larger coefficient η, applied only in late layers (L26–31 on Llama-3) where refusal is strongest: $\mathbf{W}_{v,\ell} \leftarrow \mathbf{W}_{v,\ell} + \eta\,\hat{\mathbf{u}}_\ell\hat{\mathbf{r}}_\ell^\top$. Concentrating the coupling where refusal is already causally active makes small ablation errors more consequential, because any residual refusal coordinate is immediately injected into the value stream in the final computation stages.

Head Amplification. We score heads by $s_{\ell,h} = \left\|\hat{\mathbf{r}}_\ell^\top \mathbf{W}_{o,\ell}^{(h)}\right\|_2$, select the top-$k$, and scale those head outputs by a factor γ (e.g., k=4, γ=2). If a subset of heads carries disproportionate refusal-aligned signal, amplifying them increases the magnitude of refusal-correlated residual updates, so a fixed-strength (or mis-targeted) ablation leaves more refusal behind. Because it edits $\mathbf{W}_o$ (part of Heretic's weight-edit surface), this mechanism is mostly a control for standard RFA rather than a robust defense.

Autoregressive Embedding Boost. For refusal-prefix tokens $t \in \{\texttt{I}, \texttt{cannot}, \texttt{can}, \texttt{'t}, \texttt{Sorry}\}$, we add a small refusal-aligned component to their embeddings: $\mathbf{W}_{\text{emb}}[t,:] \mathrel{+}= \alpha\,\hat{\mathbf{r}}_{\text{avg}}$, where $\hat{\mathbf{r}}_{\text{avg}}$ is the layer-averaged refusal direction. Once generation begins with a refusal prefix, the boosted embedding injects extra refusal signal at the next step, creating a positive feedback loop that can make refusal harder to interrupt if an attacker has only partially removed refusal features.

RMSNorm Gain Control. We increase RMSNorm gain on the top-k refusal-aligned channels: $\gamma_{\ell,j} \leftarrow g\,\gamma_{\ell,j}$ for indices j with the largest $|\hat{\mathbf{r}}_{\ell,j}|$. RMSNorm applies state-dependent rescaling ($\mathbf{y} = \mathbf{x}/\text{RMS}(\mathbf{x}) \odot \boldsymbol{\gamma}$), so amplifying refusal-aligned channels changes the effective coupling between channels in a data-dependent way. This could make harmful–safe differences less well described by a fixed linear direction; in practice, such edits often destabilize generation.

Orthogonal Debiasing (utility repair). We estimate an over-refusal direction $\hat{\mathbf{v}}$ from benign prompts the model incorrectly refuses (constructed to be approximately orthogonal to $\hat{\mathbf{r}}_\ell$), and remove sensitivity to $\hat{\mathbf{v}}$ by projecting it out of selected matrices: $\mathbf{W}' = \mathbf{W} - (\mathbf{W}\hat{\mathbf{v}})\hat{\mathbf{v}}^\top$ for input-dimension matrices ($\mathbf{W}_{\text{emb}}$), and $\mathbf{W}' = \mathbf{W} - \hat{\mathbf{v}}(\hat{\mathbf{v}}^\top\mathbf{W})$ for output-dimension matrices ($\mathbf{W}_o$, $\mathbf{W}_{\text{down}}$). Since DDO can mildly increase benign refusals on some models, removing the dominant benign-bias direction repairs compliance without directly weakening refusal, because the edited direction is chosen to be orthogonal to refusal.

Refusal Direction Rotation. We apply a fixed Givens rotation in the $\{\hat{\mathbf{r}}, \mathbf{v}\}$ plane ($\theta = \pi/3$, $\mathbf{v} \perp \hat{\mathbf{r}}$), folded into $\mathbf{W}_o$ and $\mathbf{W}_{\text{down}}$. This can appear to "hide" refusal by moving it away from $\hat{\mathbf{r}}$, but the rotated refusal direction $\mathbf{r}_{\text{def}} = \mathbf{M}\mathbf{r}_{\text{base}}$ is still a single linear direction, so an adaptive attacker can re-estimate DIM on the defended checkpoint and recover it (Llama-3: 68% ASR).

Representation Rerouting. We learn a rank-16 update $\Delta\mathbf{W}$ (applied to $\mathbf{W}_o$ and $\mathbf{W}_{\text{down}}$) that maps harmful activations toward random target representations. If harmful activations are pushed into a more diverse set of directions, the harmful–safe contrast could become less low-rank, weakening rank-1 DIM estimation. In practice, harmful and safe activations remain linearly separable after rerouting, so a re-estimated DIM direction still removes refusal (Llama-3: 86% ASR).

Refusal Scattering. We attempt to distribute the refusal direction across $k{=}8$ random orthogonal residual directions (via rank-k perturbations to $\mathbf{W}_o$) and boost refusal-prefix logits via LM-head scaling. If refusal were spread across multiple orthogonal directions, a rank-1 ablation would remove only a fraction of it, forcing the attacker to use higher rank or more phases. In practice, this approach performed poorly in early tests and we did not evaluate it beyond a coherence check.

LM-Head Multiplicative Row Scaling. We reshape a small set of logits by multiplicatively rescaling corresponding LM-head rows. For selected token IDs t (refusal-prefix tokens such as "I/cannot/can" and compliance tokens such as "'d/would/Sure/Here"), we apply

$$\mathbf{W}_{\text{lm}}[t,:] \leftarrow s_t\,\mathbf{W}_{\text{lm}}[t,:], \tag{10}$$

which rescales $z_t = \langle \mathbf{W}_{\text{lm}}[t,:], \mathbf{h}\rangle$ without modifying hidden-state activations. Since RFA-like attacks operate on hidden states, changing the LM head can provide additional refusal "headroom" even if intermediate refusal features are partially reduced. We tune $\{s_t\}$ with Optuna [25].

Adversarial Sign Coupling. We repurpose a small set of SwiGLU units to implement a *bilinear* cross-term between the refusal coordinate and an orthogonal coordinate. In layer ℓ, sample $\hat{\mathbf{o}}_\ell \perp \hat{\mathbf{r}}_\ell$ and choose a small set of intermediate units i disjoint from the gated-decoy units (we use a fixed neuron-index offset, 30+). We set each unit so that its gate row aligns with $\hat{\mathbf{o}}_\ell$ and its up row aligns with $\hat{\mathbf{r}}_\ell$, while its down column writes into a fixed random output direction $\mathbf{u}_{\ell,i}$. The resulting contribution is

$$\delta_i(\tilde{\mathbf{x}}_\ell) = \text{silu}(\langle \hat{\mathbf{o}}_\ell, \tilde{\mathbf{x}}_\ell\rangle) \cdot \langle \hat{\mathbf{r}}_\ell, \tilde{\mathbf{x}}_\ell\rangle \cdot \mathbf{u}_{\ell,i}. \tag{11}$$

Adaptive multi-phase RFA repeatedly estimates and ablates a *linear* DIM direction at fixed hookpoints. Because the sign-coupled term depends on the *product* of two coordinates, the effective linear surrogate presented by this bilinear feature can shift after ablation and across hookpoints, potentially increasing the number of distinct directions (phases/rank) an attacker must remove to fully suppress refusal-related computation.

E Theoretical Proofs and Formalizations

E.1 Estimator-Corruption Theorem Proof

Proof of Theorem 1. Let $U = U_k(C) \in \mathbb{R}^{d\times k}$ contain the top-k left singular vectors of C, and let $\Sigma = \text{diag}(\sigma_1(C), \ldots, \sigma_k(C))$ be the corresponding singular values. By the SVD, U satisfies

$$CC^\top U = U\Sigma^2. \tag{12}$$

Multiplying both sides by Σ^{-2} (well-defined since $\sigma_k(C) > 0$) yields $U = CC^\top U\Sigma^{-2}$. Projecting onto the causal refusal subspace $\mathcal{R}$:

$$\Pi_{\mathcal{R}} U = \Pi_{\mathcal{R}} CC^\top U\Sigma^{-2} = \Pi_{\mathcal{R}} C(C^\top U)\Sigma^{-2}. \tag{13}$$

Since $C^\top U = W\Sigma$ for an orthonormal W (from the SVD of C), we obtain

$$\Pi_{\mathcal{R}} U = \Pi_{\mathcal{R}} CW\Sigma^{-1}. \tag{14}$$

Taking operator norms and applying sub-multiplicativity:

$$\|\Pi_{\mathcal{R}} U\|_{\text{op}} \le \|\Pi_{\mathcal{R}} C\|_{\text{op}} \cdot \|W\|_{\text{op}} \cdot \|\Sigma^{-1}\|_{\text{op}} = \frac{\|\Pi_{\mathcal{R}} C\|_2}{\sigma_k(C)}, \tag{15}$$

where $\|W\|_{\text{op}} = 1$ and $\|\Sigma^{-1}\|_{\text{op}} = 1/\sigma_k(C)$. Since the attacker's ablation projector is $\Pi_k(C) = UU^\top$, we have $\|\Pi_{\mathcal{R}}\Pi_k(C)\|_{\text{op}} = \|\Pi_{\mathcal{R}} U\|_{\text{op}}$, completing the proof. □

Corollary 4 (Decoy-dominated contrast defeats rank-k attacks)**.** *If $C = C_{\text{ref}} + C_{\text{dec}} + E$ where $\Pi_{\mathcal{R}} C_{\text{dec}} = 0$ (decoys orthogonal to refusal), $\|C_{\text{ref}} + E\|_2 \le s_{\text{ref}}$, and $\sigma_k(C_{\text{dec}}) \ge s_{\text{dec}} > s_{\text{ref}}$, then:*

$$\|\Pi_{\mathcal{R}}\Pi_k(C)\|_{\text{op}} \le \frac{s_{\text{ref}}}{s_{\text{dec}} - s_{\text{ref}}}. \tag{16}$$

Proof of Corollary 4. We bound each factor in Theorem 1 separately. For the numerator: since $\Pi_{\mathcal{R}} C_{\text{dec}} = 0$ (decoy directions are orthogonal to the causal refusal subspace by construction),

$$\|\Pi_{\mathcal{R}} C\|_2 = \|\Pi_{\mathcal{R}}(C_{\text{ref}} + E)\|_2 \leq \|C_{\text{ref}} + E\|_2 \leq s_{\text{ref}}. \tag{17}$$

For the denominator: by Weyl's singular value perturbation inequality applied to $C = C_{\text{dec}} + (C_{\text{ref}} + E)$,

$$\sigma_k(C) \geq \sigma_k(C_{\text{dec}}) - \|C_{\text{ref}} + E\|_2 \geq s_{\text{dec}} - s_{\text{ref}}. \tag{18}$$

Substituting both bounds into Theorem 1 yields $\|\Pi_{\mathcal{R}} \Pi_k(C)\|_{\text{op}} \leq s_{\text{ref}}/(s_{\text{dec}} - s_{\text{ref}})$. □

Proof of Theorem 2. Since $U^\top U = I_m$, the nonzero singular values of $D_\theta = UA_\theta$ equal those of A_θ. By Weyl's inequality, $\sigma_k(C_\theta) = \sigma_k(D_\theta + S_\theta) \geq \sigma_k(D_\theta) - \|S_\theta\|_2 = \sigma_k(A_\theta) - \rho$. Because $\Pi_{\mathcal{R}} U = 0$, we have $\|\Pi_{\mathcal{R}} C_\theta\|_2 = \|\Pi_{\mathcal{R}} S_\theta\|_2 = \rho_{\mathcal{R}}$. Substituting into Theorem 1 gives the bound. The rank-budget statement follows by requiring $\rho_{\mathcal{R}}/(\sigma_k(A_\theta) - \rho) \leq \varepsilon$. □

Proof of Corollary 3. Since $\sum_j \sigma_j(A_\theta)^2 = \|A_\theta\|_F^2 \leq B^2$ and $\sigma_1 \geq \cdots \geq \sigma_k$, we have $k\sigma_k(A_\theta)^2 \leq \sum_{j=1}^{k} \sigma_j(A_\theta)^2 \leq B^2$, so $\sigma_k(A_\theta) \leq B/\sqrt{k}$. Equality holds when $\sigma_1 = \cdots = \sigma_k = B/\sqrt{k}$. □

E.2 Capability-Constrained Stackelberg Formalization

We formalize the defender–attacker interaction as a capability-constrained Stackelberg game.

Setup. The defender commits to a post-training weight edit $\theta \in \Theta$, producing checkpoint M_θ. After observing M_θ, the attacker chooses an attack $a \in \mathcal{A}$ (e.g., rank-k RFA, P-phase adaptive RFA, T-trial Heretic) to maximize harmful compliance while preserving capability:

$$u_A(\theta, a) = A(\theta, a) - \lambda[\tau - U(\theta, a)]_+ - c(a), \tag{19}$$

where $A(\theta, a) = \text{ASR}(M_{\theta,a})$, $U(\theta, a)$ is a utility proxy (e.g., MT-Bench), τ is an attacker utility threshold, λ controls the capability–compliance tradeoff, and $c(a)$ is attack cost. The defender solves

$$\theta^\star \in \arg\min_{\theta \in \Theta} \max_{a \in \text{BR}_{\lambda,\tau}(\theta)} \left[A(\theta, a) + \mu[\tau_D - U(\theta, \varnothing)]_+ + C_D(\theta)\right], \tag{20}$$

where $\text{BR}_{\lambda,\tau}(\theta) = \arg\max_{a \in \mathcal{A}} u_A(\theta, a)$ is the attacker's best response, μ penalizes defender utility violations, τ_D is the defender's utility floor (without attack), and $C_D(\theta)$ is editing cost.

Definition 5 (Capability-constrained ASR and deterrence)**.** *For defense θ, utility threshold τ, and attack family $\mathcal{A}$:*

$$\text{CC-ASR}_\tau(\theta) \triangleq \max_{a \in \mathcal{A}:\, U(\theta,a) \geq \tau} A(\theta, a). \tag{21}$$

*We say θ is (q, τ)-*deterrent *against $\mathcal{A}$ if* $\text{CC-ASR}_\tau(\theta) < q$.

CC-ASR separates clean jailbreaks (high ASR with preserved utility) from pyrrhic failures (high ASR with destroyed utility). Worst-case ASR conflates these; CC-ASR isolates the deployment-relevant threat.

Connection to the spectral toll. Under Theorem 2, if every capability-preserving attack has effective rank at most $r_{\text{eff}}(\varepsilon)$ (the estimated decoy phase budget), then $\text{CC-ASR}_\tau(\theta) \leq A(\theta, \varnothing) + L\varepsilon$, where L is an assumed Lipschitz constant relating causal overlap to a continuous surrogate of ASR (empirical ASR is discrete; we treat L as modeling a smooth expected compliance rate). DDO's objective against contrastive attackers is therefore to maximize $\sigma_k(A_\theta) - \rho$ subject to utility constraints—raising the spectral toll above the attacker's capability-preserving rank budget.

Adaptive RFA as a stopping game. For adaptive multi-phase RFA, the attacker sequentially decides whether to continue from phase k to $k{+}1$. Continuation is rational only when $\Delta A_{k+1} > \lambda\Delta[\tau - U_{k+1}]_+ + \Delta c_{k+1}$: the marginal ASR gain must exceed the marginal capability penalty plus marginal cost. DDO does not make $\Delta A_k = 0$ indefinitely; it aims to push large ASR gains to later phases where capability degradation dominates. Empirically (Figure 2), DDO (rank 8) maintains MT-Bench ≥ 5.82 through Phase 8, comparable to RepBend (6.17) and Circuit Breakers (5.90); ReFAT drops to MT-Bench 2.46 at Phase 1, so its low ASR is bought at the cost of utility under this formulation.

E.3 Spectral Theory Validation

We empirically validate the flat-spectrum tradeoff predicted by Corollary 3. We train DDO with $K \in \{1, 2, 4, 8\}$ decoy neurons per layer, each with a *diversified reader* (distinct trigger direction via random orthogonal perturbation). Without diversified readers, all K neurons read the same scalar $\langle \hat{\mathbf{r}}_\ell, \mathbf{h}_\ell \rangle$ and A_θ is effectively rank 1 regardless of K. With diversified readers, each neuron reads a different coordinate, giving A_θ up to K significant singular values. We then attack each model with rank-k SVD ablation (Table 6).

Table 6: Diversified-reader DDO under rank-k attacks (ASR %↓, LlamaGuard-2, Llama-3-8B). K = decoy neurons per layer; k = attack rank. As predicted by Corollary 3, increasing K weakens rank-1 protection (energy spread) but strengthens higher-rank protection (more decoy directions to exhaust).

Attack rank k	K=1	K=2	K=4	K=8
k=1	10	21	27	14
k=2	1	1	0	0
k=4	1	0	0	1
k=8	2	1	1	8
k=16	5	2	2	2

Three patterns confirm the corollary: (i) *rank-1 protection weakens with K*: ASR rises from 10% (K=1) to 27% (K=4) as energy is spread across more directions; (ii) *higher-rank protection strengthens*: at k=16, ASR drops from 5% (K=1) to 2% ($K \geq 2$); and (iii) K=2–4 provides the best tradeoff (0–1% ASR at k=2–8 with moderate rank-1 cost).

Empirical singular spectrum. Table 7 reports the singular values of A_θ (the empirical decoy response matrix from Theorem 2) for three configurations.

Table 7: Singular values of the decoy response matrix A_θ on Llama-3-8B DDO. Diversified readers flatten the spectrum (σ_1/σ_2 drops from 934:1 to 15:1), confirming the mechanism underlying Corollary 3.

Configuration	σ_1	σ_2	σ_3	σ_4	σ_1/σ_2
K=1, single reader	158.8	0.17	0.16	0.04	934:1
K=4, single reader	339.0	2.33	1.66	1.05	145:1
K=4, diversified reader	82.6	5.59	1.55	1.15	**15:1**

Diversified readers reduce σ_1/σ_2 from 934:1 (single reader, effective rank 1) to 15:1 (diversified, effective rank >1), directly confirming that reader diversification flattens the decoy spectrum as predicted.

F Full Experimental Details

F.1 Attack Implementation Details

Standard RFA. We compute per-layer DIM directions at the last token position: for each layer ℓ, $\hat{\mathbf{r}}_\ell = \text{normalize}(\bar{\mathbf{h}}_\ell^{\text{harm}} - \bar{\mathbf{h}}_\ell^{\text{safe}})$ where $\bar{\mathbf{h}}$ denotes the mean activation over the probe set. Mean activations are accumulated in float64 for numerical stability. All 32 per-layer directions are applied simultaneously via forward hooks (three-point: block input + attention output + MLP output; residual-stream: block input only). This differs from the Arditi et al. direction *selection* pipeline [7], which selects a single (position, layer) direction by evaluating each candidate using KL divergence on safe prompts. Our per-layer procedure corresponds to a larger ablation budget (one direction per layer) and is therefore at least as strong as the single-direction variant under matched probe sets. For all reported results we use **unfiltered** DIM estimation (no refusal-score filtering of the probe set), which provides the attacker with a cleaner direction estimate.

Adaptive multi-phase RFA. At each phase k, we collect activations with all prior phases' hooks active, compute fresh per-layer DIM directions, and Gram-Schmidt orthogonalize against all previous directions. All k direction sets are then ablated simultaneously using three-point hooks. Total hookpoints: up to $3kL$ per forward pass (3 hookpoints $\times$ k phases $\times$ L=32 layers). The attack is deterministic: three identical runs produce identical direction sequences. We evaluate up to 8 phases.

Heretic [14]. We run the released Heretic tool in an isolated Python virtual environment via subprocess. Key settings: Optuna TPE sampler with `n_startup_trials=15`, `n_ei_candidates=128`, `multivariate=True`. Multi-objective optimization: minimize (refusals, KL divergence). The tool optimizes a peaked per-layer ablation strength profile $\alpha(\ell)$ over $\mathbf{W}_o$ and $\mathbf{W}_{\text{down}}$ matrices. We report results at 200 trials (H-200). Direction scope is searched over "global" (single direction index) and "per_layer". Per-component search ranges: `max_weight` $\in [0.8, 1.5]$, `max_weight_position` $\in [0.6L, L]$, `min_weight` $\in [0, 1] \times$ `max_weight`, `min_weight_distance` $\in [1, 0.6L]$.

F.2 DDO and Auxiliary Mechanism Hyperparameters (Llama-3)

Table 8 reports exact hyperparameters for DDO and all auxiliary mechanisms on Llama-3. All mechanisms share the same unfiltered DIM estimation protocol and probe budget.

Table 8: Hyperparameters for the 15 mechanisms used in reported configurations on Llama-3-8B-Instruct. Shared: bfloat16 weights, float64 direction estimation, 128 harmful + 128 safe probes, unfiltered DIM, last-token position. DDO (first row) is the core mechanism.

Mechanism	Scope	Key parameters
Core mechanism:		
DDO (Gated Decoy)	varies	1 neuron/layer (low_norm). Orthogonal decoy basis $\mathbf{u}_i \perp \hat{\mathbf{r}}_\ell$, gradient-optimized. Per-model layer ranges and hyperparameters in Table 13. Modifies $\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{up}}, \mathbf{W}_{\text{down}}$
Auxiliary mechanisms (examples; used in some compositions/ablations):		
Value-Projection Conditioning	12–19	Rank-1 perturbation of $\mathbf{W}_v$ (random output direction)
Head Amplification	24–28	Top 4 heads, 2× factor
Orthogonal Debiasing	all	$\mathbf{W}' = \mathbf{W} - (\mathbf{W}\hat{\mathbf{v}})\hat{\mathbf{v}}^\top$ on $\mathbf{W}_{\text{emb}}, \mathbf{W}_o, \mathbf{W}_{\text{down}}$
Standalone variants:		
LM-Head Multiplicative Row Scaling	$\mathbf{W}_{\text{lm}}$	Optuna TPE, 40 trials, 9 tokens (e.g. "I": 2.63×, "Sure": 0.32×)
Decoy Shear Transform	all	$\mathbf{M} = \mathbf{I} + \mathbf{UTU}^\top$; sweep $c \in \{1-5\}$, $k \in \{4, 8, 16\}$
Refusal Scattering	varies	Rank-k scatter (k=8) + LM-head boost
Representation Rerouting	varies	Rank-16 $\Delta\mathbf{W}$, $\lambda_{\text{reroute}}/\lambda_{\text{preserve}}$
Refusal Direction Rotation	all	Givens in $\{\hat{\mathbf{r}}, \mathbf{v}\}$ plane, $\theta = \pi/3$
Adaptive-specific additions (used in adaptive compositions):		
Diverse Q-Head Routing	16–30	16–22 heads, q_{strength} 0.59–0.75, per-head rotated triggers
RoPE Dimension Amplification	16–28	1.45–1.54× on 11–16 rotary dims
Autoregressive Embedding Boost	$\mathbf{W}_{\text{emb}}$	Strength 0.68–0.82, tokens {I, cannot, can, 't, Sorry}
KV Rank-1 Strengthening	26–31	Strength 0.55–0.60, rank-1 on $\mathbf{W}_v$
RMSNorm Gain Control	16–28	1.07–1.30× on 59–64 refusal-aligned γ channels
Adversarial Sign Coupling	4–27	Gate on $\hat{\mathbf{o}} \perp \hat{\mathbf{r}}$, up on $\hat{\mathbf{r}}$, neuron offsets 30+

F.3 Per-Model Causal Refusal Zones

To choose DDO layers, we estimate each model's causal refusal zone via single-layer ablations. For each layer ℓ, we ablate only the per-layer DIM direction at ℓ (3-point hooks) and measure the resulting drop in refusal rate (substring heuristic) on 16 harmful prompts. We classify a layer as "causal" if the refusal drop exceeds 10%. Table 9 reports the resulting causal zones and DDO placement for all 7 models. "Causal Zone" is the span from the earliest to the latest causal layer; #C is the number of layers within that span that exceed the 10% threshold (not necessarily contiguous—e.g., Llama-2 has 3 causal layers scattered across L13–31); and Frac is #C divided by the total number of layers. The upstream placement principle (Section 3) suggests that, on localized models, DDO layers should start at or before the causal zone (allowing partial overlap).

Table 9: Causal refusal zones, DDO placement, and compile mode across all 7 models. #C = number of causal layers; Frac = fraction of total layers that are causal. Compile mode does not follow a simple redundancy rule: Mistral has 81% causal fraction (distributed) yet requires additive because replace inverts its refusal signal (Table 12).

Model	Causal Zone	#C	Frac	Pattern	DDO Layers	Mode
Yi (48L)	L3–45	42	88%	Distributed	L18–32	Rep
Llama-2 (32L)	L13–31	3	9%	Sparse	L4–12	Rep
Llama-3 (32L)	L8–31	8	25%	Localized	L6–14	Rep
Gemma-2 (42L)	L12–23	12	29%	Localized	L10–18	Add
Qwen3 (36L)	L20–34	15	42%	Localized	L12–16	Add
Mistral (32L)	L2–28	26	81%	Distributed	L2–18	Add
GLM-4 (40L)	L16–22	6	15%	Localized	L6–17	Rep

F.4 Compile Mode Selection Details

We compile DDO parameters into model weights using one of two modes: (i) **replace** (`.copy_()`), which completely overwrites the neuron's row/column vectors with the decoy trigger and output (a clean, high-magnitude decoy signal, but it removes the neuron's original computation), and (ii) **additive** (`.add_()`), which adds the decoy signal on top of the original neuron weights (preserves original behavior, but yields a weaker, mixed decoy signal).

The preferred mode depends on the model's **refusal redundancy**—whether refusal is *distributed* across many layers (ablating any single layer only partially reduces refusal) or *localized* in a narrow band (ablating any one causal layer causes near-complete refusal loss):

Table 10: DDO compile mode comparison across 7 models (ASR %↓, mean of 4 RFA variants, 3-judge avg, 15 Optuna trials per mode). best / poor / ours; × = coherence failure. The preferred mode is model-specific; replace is stronger when tolerated but breaks some architectures.

Model	Gate	Replace↓	Additive↓	Best
Yi-1.5-9B	SwiGLU	**5%**	18%	Replace
Llama-2-7B	SwiGLU	**1%**	8%	Replace
Llama-3-8B	SwiGLU	**0%**	7%	Replace
Gemma-2-9B	GeGLU	×	**0%**	Additive
Qwen3-8B	SwiGLU	18%	**8%**	Additive
Mistral-7B	SwiGLU	42%	**0%**	Additive
GLM-4-9B	Fused	**2%**	48%	Replace

Replace mode produces a stronger decoy signal but is more disruptive: overwriting a neuron's weights removes its original computation. On some models this eliminates baseline refusal entirely (Mistral: 0% Ref under replace; Table 12), while on others the model compensates (Yi: 100% Ref despite replacement). Gemma-2 (GeGLU) fails the coherence check under replace entirely. Additive mode preserves the original neuron and superposes the decoy, making it safer but producing a weaker, mixed decoy signal. Note that the closed-form decoy contribution (Equation 1) is exact only under replace mode; in additive mode, the SwiGLU nonlinearity produces cross-terms between the original and inserted weights, so Equation 1 should be interpreted as the intended decoy component rather than the exact neuron output. In practice, we evaluate both modes and select by ASR subject to passing the coherence check.

F.5 Layer Placement Analysis

DDO's effectiveness depends critically on *where* decoys are placed relative to the model's causal refusal zone—the layers whose ablation causally reduces refusal (Table 9). Ideally, decoy layers should start at or before the causal zone so that the decoy signal contaminates the attacker's estimator without disrupting the refusal computation itself.

Table 11 shows that placement affects baseline refusal: placing decoys *inside* the causal zone can substantially reduce refusal (e.g., Gemma-2: 93% → 68%; Qwen3: 37% → 19%; Mistral:

Table 11: Effect of decoy placement on baseline refusal rate. Refusal rate measured on 100 JailbreakBench behaviors (substring heuristic, without attack). DDO uses per-model Optuna-tuned hyperparams (Table 13). Ref_{base} = undefended model; Ref_{in} = decoys placed inside the causal zone; Ref_{up} = decoys placed upstream.

Model	Causal	$\text{Ref}_{\text{base}} \uparrow$	$\text{Ref}_{\text{in}} \uparrow$	$\text{Ref}_{\text{up}} \uparrow$	Mode
Yi	L3–45	75%	93%	93%	Replace
Llama-2	L13–31	96%	94%	96%	Replace
Llama-3	L8–31	95%	91%	95%	Replace
Gemma-2	L12–23	93%	68%	93%	Additive
Qwen3	L20–34	37%	19%	35%	Additive
Mistral	L2–28	58%	3%	18%	Additive
GLM-4	L16–38	96%	72%	96%	Replace

$58\% \rightarrow 3\%$), while *upstream* placement largely preserves it. On models with distributed refusal (Yi, Llama-2, Llama-3), placement makes less difference because refusal is redundant across layers. Placement also affects *ASR under attack*: on Qwen3, decoys placed inside the causal zone yield 80% 3-point RFA ASR (LlamaGuard-2, JailbreakBench) compared to 35% for upstream placement, even though baseline refusal is preserved in both cases. This motivates the upstream placement principle: placing decoys at or before the causal zone contaminates the attacker's estimator without disrupting the refusal computation itself.

Table 12 shows how compile mode affects baseline refusal:

Table 12: Baseline refusal rate (↑) under DDO replace vs. additive. Refusal rate measured on 100 JailbreakBench behaviors (substring heuristic, without attack). DDO uses per-model Optuna-tuned hyperparams (Table 13). red = refusal disrupted by DDO. The best compile mode depends on how each model's refusal circuitry responds to neuron overwriting.

Model	$\text{Ref}_{\text{rep}} \uparrow$	$\text{Ref}_{\text{add}} \uparrow$	Best
Yi	93%	56%	Rep
Llama-2	96%	94%	Rep
Llama-3	95%	78%	Rep
Gemma-2	74%	8%	Add
Qwen3	52%	31%	Add
Mistral	3%	18%	Add
GLM-4	96%	38%	Rep

Table 12 reveals why the best compile mode is model-specific: replace mode produces a stronger decoy but can disrupt refusal itself. On Mistral, replace nearly eliminates refusal (3% Ref), explaining its higher ASR with replace vs. additive (Table 10). On GLM-4, the pattern reverses: additive disrupts refusal (38% Ref) while replace preserves it (96%). The best mode therefore depends on how the model's refusal circuitry responds to neuron overwriting, not on a simple rule.

F.6 Per-Model DDO Best Configs

Table 13 reports the DDO hyperparameters used for each model in Table 3. For each model, we run 15 Optuna trials in replace mode and 15 in additive mode, selecting the lower-ASR mode subject to a benign-compliance drop of < 20 percentage points on held-out validation prompts when feasible. The search space covers layer range, β (gate scale), scale (output magnitude), confusion λ (estimator-confusion loss weight), learning rate, and number of epochs. Layer ranges are guided by the causal refusal zone (Table 9): on localized models we constrain DDO layers to start at or before the causal zone; on distributed models the search is unconstrained.

F.7 Benchmark Settings

Decoding and coherence. Unless stated otherwise, we use greedy decoding (`do_sample`=false) with `max_new_tokens`=512 and batch size 8; all timing measurements use a single A100 GPU. Some post-hoc edits can break generation (e.g., near-empty or highly repetitive outputs) even when

Table 13: DDO hyperparameters per model (Optuna-tuned, 15 trials per compile mode). Replace-mode params shown for Yi/Llama-2/Llama-3/GLM-4; additive-mode params for Gemma-2/Qwen3/Mistral. lr = learning rate, ep = epochs. DDO generalizes across architectures with per-model Optuna tuning of 6 hyperparameters.

Model	Mode	Layers	β	scale	conf.λ	lr	ep
Yi	Replace	L18–32	6.1	0.42	0.5	0.002	2
Llama-2	Replace	L4–12	2.9	0.47	1.1	0.005	3
Llama-3	Replace	L6–14	1.30	0.46	1.12	0.003	2
Gemma-2	Additive	L10–18	6.42	0.36	1.79	0.021	1
Qwen3	Additive	L12–16	3.20	0.08	1.79	0.012	3
Mistral	Additive	L2–18	4.05	0.46	0.82	0.008	2
GLM-4	Replace	L6–17	5.3	0.29	1.2	0.016	1

ASR appears low, so we apply a deterministic coherence check on 10 benign prompts and drop configurations with empty, degenerate (single-token repetition $>80\%$), or too-short outputs (marked "$\times$").

Splits and DIM estimation. Hyperparameter tuning and DIM estimation use a held-out validation split (128 harmful + 128 safe probes); benchmark metrics are reported on held-out test sets. For DDO optimization, DIM directions are computed per-layer from last-token post-attention-layernorm activations ($\tilde{\mathbf{x}}_\ell$, the MLP input) using **unfiltered** estimation (no refusal-score filtering), accumulated in float64. For attack evaluation (RFA), directions are estimated from residual-stream activations ($\mathbf{h}_\ell$), matching the standard attacker protocol.

Table coloring conventions. Throughout all tables, cell shading follows consistent thresholds relative to the per-block base model. For ASR-type metrics (DR, RFA, Heretic, GCG, PAIR, AutoDAN; lower is better): **best** = best in column; good = $\leq 10\%$; poor = $\geq 50\%$; unshaded = 10–50%. For XSTest (higher is better): good = within 5pp of base or improved; poor = drop >20pp. For MMLU (higher is better): good = within 1pp of base or improved; poor = drop >3pp. For MT-Bench (higher is better): good = within 0.5 of base or improved; poor = drop >2. ours = DDO rows.

MMLU. We evaluate 5-shot multiple-choice accuracy on MMLU's 57 subjects. We run evaluation via `lighteval` using the `helm|mmlu:<subject>|5|1` task configuration (5-shot; few-shot truncation enabled) and report average exact-match accuracy across subjects. For tractability during iterative defense development, we cap evaluation to 100 questions per subject unless otherwise noted.

MT-Bench. We run the standard MT-Bench benchmark (80 multi-turn questions) using FastChat's `llm_judge` pipeline with GPT-4 as the judge in single-mode and report the average score (1–10) across questions.

XSTest. We measure over-refusal on XSTest [38]—250 safe prompts that superficially resemble harmful ones (e.g., "How do I kill a process in Linux?"). A defended model is expected to comply; refusal here is a false-positive cost of the defense. We generate one response per prompt with greedy decoding (`do_sample`=False, `max_new_tokens`=200), then classify each response with GPT-4o under the official XSTest 3-class rubric: *full compliance* (direct answer), *full refusal* (declines or only offers alternatives), and *partial refusal* (mixed response). The reported XSTest score is the share of full-compliance outputs; partial and full refusal both count against the model.

F.8 Baseline Sweep Details

Common training setup. We trained six defenses (Circuit Breakers, RepBend, Triplet, Triplet-Adv, LAT, ReFAT) on six model families (Llama-3, Yi-1.5, Qwen3, Gemma-2, Mistral, GLM-4). All methods use LoRA fine-tuning on `q,k,v,o,gate_proj,up_proj,down_proj` (GLM-4 uses `query_key_value, dense, dense_h_to_4h, dense_4h_to_h`), bf16 precision with tf32 matmuls, a constant learning-rate scheduler, gradient checkpointing, 1 epoch, and an effective batch size of 16 (`per_device_train_batch_size=2`, `gradient_accumulation_steps=8`). All CRL-family methods (CB, RepBend, Triplet, Triplet-Adv) train on `allenai/wildguardmix`

(`wildguardtrain` split). For each (defense, model) pair, we run a 3-config sweep (A/B/C) over the primary hyperparameter(s) identified in the source paper; the best config is selected by lowest RFA ASR averaged over three judges (HarmBench classifier, LlamaGuard-2, StrongREJECT).

Circuit Breakers [9]. Circuit Breakers fine-tunes a LoRA adapter with a representation-rerouting loss that pushes harmful-prompt residual activations toward a "broken" (orthogonal or random) target representation, while a retention term preserves benign activations. We sweep `loss_alpha` (rerouting strength) over $\{10, 20, 40\}$, with all other loss terms set to zero (`loss_beta=loss_gamma=loss_epsilon=loss_eta=0`). Training uses LR 10^{-4}, LoRA r=16, α=16, dropout 0.05, `max_seq_length=2048`, and 150 steps (Config D: 300 steps). Target layers follow the last 30–50% of each model (Table 14).

RepBend [12]. RepBend extends CB with a representation-bending loss: additional penalty terms (β, γ, ε) reshape the geometry of harmful representations along a learned direction while preserving benign ones. We jointly sweep (`loss_alpha, loss_gamma`) over $\{(0.5, 0.3), (0.75, 0.45), (1.0, 0.6)\}$, fixing `loss_beta=0.1`, `loss_epsilon=0.3`, `loss_mode=response_all`, `alpha_mode=all`. Training uses LR 10^{-5}, LoRA r=16, `max_seq_length=4096`, 450 steps (D: 900 steps).

Triplet [13]. The Triplet loss trains the model to pull harmful-prompt activations toward refusal anchors and push them away from compliance anchors in representation space, using margins (`margin_p, margin_n`). We sweep over $\{(2, 3), (4, 6), (8, 12)\}$. Fixed args: `loss_alpha=0.5`, `loss_beta=0.6`, `loss_gamma=0.7`, `loss_epsilon=0.7`, `loss_safe_dist_p=norm`, distance metrics for unsafe pairs use cosine. Training uses LR 10^{-4}, LoRA r=16, 900 steps (D: 1800 steps). Triplet is evaluated on Llama-3 in the main comparison (Table 2); it is included in the sweep for completeness.

Triplet-Adv [13]. Identical to Triplet (`loss_attack=False` $\rightarrow$ `True`): an adversarial latent perturbation is injected into hidden states during training before the triplet loss is applied, sharpening the representation boundary against activation-space attacks. Same sweep grid and hyperparameters as Triplet.

LAT [10]. Latent Adversarial Training alternates between a PGD inner loop that perturbs hidden activations subject to an ℓ_2 budget `epsilon` and an outer loop that minimizes task loss on the perturbed activations. We sweep (`epsilon, pgd_iters`) over $\{(4, 16), (8, 24), (12, 32)\}$. Other args: `outer_lr=8e-5`, `inner_lr=1e-3`, `batch_size=16`, `max_batch_per_acc=2`, LoRA r=64, 150 steps (D: 300–400 steps).

ReFAT [11]. Refusal-Feature Adversarial Training augments each training step with probability `p_rfa`: it estimates a refusal direction via DIM over 32 samples, ablates it from activations across 75% of layers, and computes the loss on the ablated activations; the refusal direction is recomputed every 4 steps. We sweep `p_rfa` over $\{0.5, 0.7, 0.9\}$. Other args: LR 2×10^{-5}, `max_seq_length=512`, `lora_alpha=32`, gradient clip 1.0. Per-model LoRA rank and batch size: Llama-3 (r=128, batch 32); Gemma-2 (r=64, batch 8); all others (r=128, batch 8).

Layer targeting. CB and RepBend/Triplet apply their losses only to a model-specific subset of layers; LAT and ReFAT similarly focus perturbations on a fraction of the network. Table 14 lists the target-layer windows used across all six models.

Table 14: Per-model layer targeting for CB and RepBend/Triplet. CB applies the rerouting loss to the listed layer range; RepBend/Triplet use a sliding window starting at `RB_START` of width `RB_WIN`. All windows cover the last 30–50% of model depth.

Model	Total layers	CB target layers	`RB_START`	`RB_WIN`
Llama-3-8B	32	20–31	20	11
Yi-1.5-9B	48	28–47	30	16
Qwen3-8B	36	21–35	22	12
Gemma-2-9B	42	25–41	26	14
Mistral-7B	32	19–31	20	11
GLM-4-9B	40	24–39	25	13

D-config protocol. If the best A/B/C config for a (defense, model) pair still exceeds 50% RFA ASR, we retrain it at $2\times$ the step count (Config D) with all other hyperparameters fixed. Config D provided modest or no improvement in most cases; diminishing returns were common, particularly for CB, RepBend, and Triplet on models with diffuse refusal. Step counts for D: CB 300; RepBend 900; Triplet/Triplet-Adv 1800; LAT 300–400; ReFAT 300.

Config E: LLM-assisted hyperparameter selection. As an additional fairness check, we ran a Config E in which Claude Opus 4.6 proposed hyperparameters for each (defense, model) pair, conditioned on the source paper, the search ranges, and the A–D outcomes. Config E did not improve over the swept best for any pair, and did not outperform DDO under standard or adaptive RFA. Throughout the paper we report the best of $\{$A,B,C,D,E$\}$ per (defense, model) pair, so the comparison systematically favors the trained baselines over DDO.

Reproducibility caveats. LAT and ReFAT use a separate Python environment from the CRL-family methods due to dependency conflicts (`LLaMA-Factory` vs. `refusal_direction_defense`). All runs use `CUBLAS_WORKSPACE_CONFIG=:16:8` for determinism and disable Weights & Biases logging. Adapter checkpoints are saved at `baselines/<model>/<defense>_<config>/` and merged on demand for evaluation.

F.9 Compute Budget

Table 15 summarizes approximate wall-clock costs for key optimization and evaluation steps (single A100 GPU). Post-hoc editing and light tuning complete in minutes to tens of minutes (even with Optuna), enabling broad design-space sweeps; most wall time comes from *stress-testing* (Heretic and per-phase MT-Bench), i.e., evaluation dominates editing.

Table 15: Wall-clock compute on a single A100 GPU. Defense editing completes in minutes; evaluation (especially Heretic) dominates wall-clock cost.

	Step	Prompts	Time
Defense	DDO (per run)	256 (128H+128S)	∼2 min
	LM-head scaling (per run)	20	∼5 sec
Evaluation	RFA (4 variants)	100–159	∼5 min
	Heretic (200 trials)	100	∼30–60 min
	MT-Bench (per phase)	80	∼20 min

F.10 Adaptive Attack: Exact ASR and MT-Bench by Phase

Table 16 summarizes per-phase ASR (phases 1–8) for the full 8-phase setting; Table 17 reports the exact per-phase ASR (LlamaGuard-2) *and* MT-Bench (GPT-4 judge) values underlying Figure 2. Figure 4 provides a compact ASR–utility summary of the same runs.

Attack cost model. Each adaptive phase requires one DIM re-estimation (256 forward passes on 128 harmful + 128 safe probes, $\sim$30 s) followed by Gram-Schmidt orthogonalization, hook installation, and evaluation ($\sim$1–2 min), totaling $\sim$2 min per phase on a single A100 GPU. A full 8-phase attack therefore costs $\sim$15–20 min wall-clock. The attack is fully automated (a deterministic loop over phases), requires no ML expertise, and uses only the defended checkpoint and a generic probe set—the same probes used for standard RFA suffice. However, multi-phase ablation degrades the resulting checkpoint: DDO-defended Llama-3 drops from MT-Bench 7.67 (unattacked) to 6.17 at Phase 5 (peak 65% ASR) and 5.82 by Phase 8 (Table 17)—a $\sim$1.9-point utility cost. By contrast, on the undefended base model the attacker obtains 73% ASR at Phase 1 with only a 0.3-point MT-Bench drop. DDO therefore forces the attacker to trade substantially more model quality per unit of ASR gained.

These phase trajectories reinforce the main-text comparison: DDO (rank 8) preserves utility (MT-Bench ≥ 5.82 through Phase 8) at a level comparable to RepBend (6.17) and Circuit Breakers (5.90), while ReFAT exhibits a sharp utility drop early (Phase 1 MTB 2.46) despite low ASR.

Table 16: Adaptive multi-phase RFA profiles (ASR %, LlamaGuard-2) across all 8 phases. "Worst" is the maximum ASR across phases (the attacker can stop at any phase). Both trained and post-hoc defenses degrade at later phases. DDO (rank 8) achieves 65% worst-case, comparable to trained RepBend (58%), without training.

Defense	Ph1	Ph2	Ph3	Ph4	Ph5	Ph6	Ph7	Ph8	Worst↓
Base (undefended)	73	64	47	34	63	69	69	59	73
RepBend (trained)	4	20	21	28	43	56	58	42	58
Circuit Breakers (trained)	0	32	75	72	69	66	63	65	75
ReFAT (trained)	3	1	1	3	5	9	60	46	60
DDO (rank 8)	0	10	28	29	65	56	49	53	65
DDO (rank 8, Optuna)	0	11	31	29	52	46	66	75	75

Table 17: Per-phase ASR (↓) and MT-Bench (↑) under adaptive multi-phase RFA. Phase 0 = unattacked model. ASR: LlamaGuard-2; MT-B: GPT-4 judge. DDO (rank 8) maintains MT-B > 5.8 through all 8 phases, comparable to RepBend and Circuit Breakers; ReFAT drops to 2.5 at Phase 1 despite low ASR.

	Base		RepBend		Circuit Breakers		ReFAT		DDO (rank 8)		DDO (rank 8, Optuna)	
Ph	ASR	MTB	ASR	MTB	ASR	MTB	ASR	MTB	ASR	MTB	ASR	MTB
0	2	7.89	0	7.69	32	7.73	0	7.37	6	7.67	1	7.60
1	73	7.47	4	4.00	0	5.40	3	2.46	0	7.96	0	7.83
2	64	7.23	20	8.29	32	7.19	1	2.72	10	7.36	11	7.27
3	47	6.79	21	8.20	75	7.11	1	3.17	28	7.09	31	7.14
4	34	6.40	28	7.82	72	7.09	3	3.66	29	6.98	29	6.82
5	63	6.21	43	7.33	69	6.53	5	4.40	65	6.17	52	6.48
6	69	3.91	56	6.72	66	6.49	9	3.62	56	5.91	46	6.14
7	69	6.16	58	6.62	63	6.29	60	3.46	49	6.03	66	5.27
8	59	6.02	42	6.17	65	5.90	46	4.01	53	5.82	75	2.95

F.11 Extended Model Comparison

Table 18 reports additional defense configurations evaluated on Llama-3, using consistent naming: **DDO** = the core mechanism alone; **DDO + debiasing** = DDO composed with orthogonal debiasing; **DDO (rank k)** = DDO with k diversified decoy readers (for adaptive RFA). The shown compositions are selected from a broader mechanism-subset search guided by two criteria: (i) minimize RFA ASR under the target attacker tier, and (ii) preserve utility (benign compliance $> 85\%$, MT-Bench > 6.0). The rank-8 variants diversify readers to increase the decoy spectral toll under adaptive multi-phase RFA; the Heretic-targeted variant allocates mechanisms to matrices outside Heretic's weight-edit surface.

F.12 Cross-Model Mechanism Comparison and Negative Results

Table 19 compares 8 systematically evaluated mechanisms across all 7 models (10–30 Optuna trials each). Three patterns emerge: (i) DDO is the only mechanism that achieves <10% ASR on every architecture; (ii) five mechanisms (diverse Q-head routing, rotation, V-projection, head amplification, KV strengthening) break generation entirely on Yi, Gemma-2, and GLM-4 ($\times$ = coherence failure), suggesting these architectures are less tolerant of attention- and normalization-space edits; and (iii) random orthogonal decoys without gradient optimization achieve low ASR only on Yi (2%) but fail on most other models (18–53%), confirming that the optimization in DDO's 4-part loss is essential. Formal definitions are in Appendix D.2.

Negative results and failed approaches. The broader mechanism search also yielded several informative negative results, organized by failure mode (ASR: 3-point RFA, JailbreakBench, LlamaGuard-2 unless noted).

Linear edits are fragile under adaptive re-estimation. Mechanisms equivalent to an input-independent linear transformation of the residual stream are defeated once the attacker re-estimates

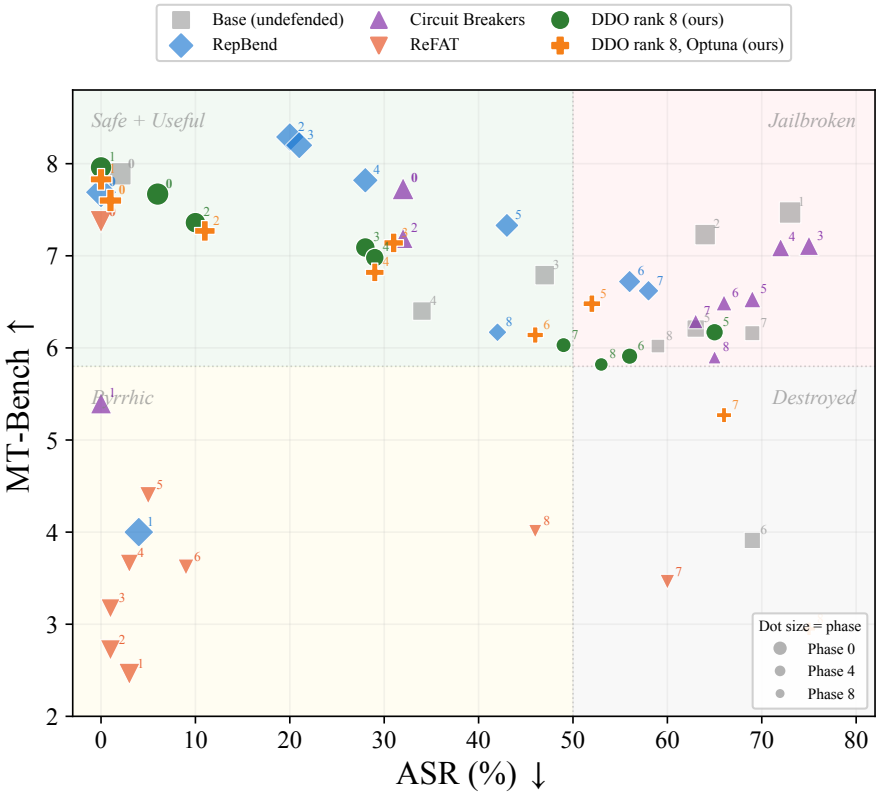

Figure 4: ASR vs. MT-Bench trajectories under adaptive attack. Each dot is one defense at one phase (0–8); larger dots = earlier phases. Quadrants: **top-left** = safe + useful (ideal); **top-right** = jailbroken but coherent (most dangerous); **bottom-left** = refuses but unusable; **bottom-right** = destroyed. ReFAT collapses into the bottom-left region (low ASR but utility destroyed), while DDO compositions (green, orange) preserve utility comparably to RepBend and Circuit Breakers, with worst-case ASR rising under sustained re-estimation.

DIM on the defended checkpoint. *Decoy Shear Transform* (92% ASR) injects orthogonal decoy components via a linear shear $\mathbf{M} = \mathbf{I} + \sum c_j \mathbf{u}_j \hat{\mathbf{r}}^\top$, but the defended DIM direction simply rotates to track the sheared mean-difference. *Refusal Direction Rotation* (68% ASR) applies a Givens rotation in the $\{\hat{\mathbf{r}}, \mathbf{v}\}$ plane; the attacker recovers the rotated direction. *Representation Rerouting* (86% ASR) maps harmful activations to random targets via a low-rank $\Delta\mathbf{W}$, but provides insufficient disruption of linear separability.

Spreading decoy budget can be counterproductive. Distributing decoy energy across more directions involves a fundamental tradeoff (Corollary 3): with identical readers, adding more output directions dilutes each one without increasing the effective rank of A_θ. With diversified readers, spreading energy trades rank-1 protection for higher-rank robustness (see Appendix E.3 for empirical validation).

Aggressive edits destroy capability. *LM-Head Row Scaling* achieves near-zero RFA ASR (0.7%) but at catastrophic utility cost (XSTest 8.0%, MT-Bench 1.72)—the model over-refuses almost all prompts. Extending DDO injection to layers 0–11 (48% ASR) degraded generation quality to incoherent outputs. These cases illustrate that low ASR is insufficient without utility preservation.

Mismatched threat models. *Gradient landscape roughening* (84% ASR) perturbs $\mathbf{W}_{\text{gate}}$ to create rough input-space loss surfaces, targeting GCG-style prompt optimization. This does not improve robustness against representation-space abliteration (RFA), confirming that input-space and representation-space attacks require distinct defensive mechanisms.

Table 18: Extended comparison on Llama-3-8B-Instruct. Utility: MMLU (%), MT-B (MT-Bench), XST (XSTest) (%). Robustness: RFA avg ASR (%), Heretic ASR (%) at 200 Optuna trials, prompt jailbreak ASR (%, 3-judge avg). † = best Heretic trial had KL > 0.3 (model degradation, not clean jailbreak). **best** / good / poor / ours. DDO alone achieves 1.8% RFA and 18% Heretic; debiasing repairs XSTest; rank-8 variants target adaptive RFA.

Defense	Type	MMLU↑	MT-B↑	XST↑	RFA↓	H-200↓	GCG↓	PAIR↓	AutoDAN↓
Base	–	**68.1**	**7.89**	94.8	85.1	88.7	27.7	55.1	0.4
Circuit Breakers	Trained	67.6	7.73	95.6	1.1	24.0	3.8	32.1	5.7
LAT	Trained	67.8	7.52	20.8	6.8	82.3	8.0	34.6	**0.0**
ReFAT	Trained	67.2	7.37	60.4	4.4	85.3	30.6	**30.0**	**0.0**
RepBend	Trained	64.4	7.69	98.4	4.8	**1.3**	0.6	38.8	0.2
Triplet	Trained	67.7	7.78	98.0	18.1	0.0†	**0.4**	39.8	**0.0**
Triplet-Adv	Trained	65.6	7.47	96.8	7.7	30.0†	22.2	43.6	20.8
DDO	Ours	66.8	7.38	91.6	1.8	18.0	1.7	37.5	**0.0**
DDO + debiasing	Ours	66.8	7.56	**99.2**	1.0	22.3	**0.4**	36.3	**0.0**
DDO with diversified readers (for adaptive RFA; Section 4.2):									
DDO (rank 8)	Ours	64.7	7.56	94.0	**0.1**	22.7†	3.4	37.9	**0.0**
DDO (rank 8, Optuna)	Ours	65.1	7.68	94.8	0.6	19.3†	3.1	39.2	0.2
Additional compositions (design-space exploration):									
DDO + 8 diversified aux[a]	Ours	64.6	7.27	94.8	0.2	**16.3**†	2.5	38.4	**0.0**
DDO + Heretic-targeted[b]	Ours	66.8	7.53	93.6	0.6	58.7	1.5	34.2	**0.0**

[a]DDO + debiasing + diverse Q-head routing + RoPE + embed boost + KV strengthening + RMSNorm + sign coupling. [b]DDO + debiasing with mechanisms in Heretic-resistant matrices.

Table 19: Cross-model mechanism comparison (ASR %↓, Optuna-tuned, 10–30 trials each). DDO row: mean of 4 RFA variants, 3-judge avg (consistent with Tables 2–3). Other mechanisms: 3-point RFA, LlamaGuard-2 (faster diagnostic). best / poor / ours; × = coherence failure. DDO is the only mechanism achieving <10% ASR across all 7 architectures.

Mechanism	Yi	Llama-2	Llama-3	Gemma-2	Qwen3	Mistral	GLM-4
DDO	5.0	**1.0**	**1.8**	**0.0**	**8.0**	**0.0**	**2.0**
Random decoy (no opt)	**2**	18	42	28	10	43	53
diverse_q	×	44	65	×	66	32	×
gate_boost	66	56	77	56	80	67	81
rotation	×	56	77	×	79	63	×
v_proj	×	54	65	×	80	69	×
head_amp	×	58	84	×	84	66	×
kv_strength	×	67	82	×	81	72	×

Random decoy uses random orthogonal directions (no gradient optimization).

F.13 Reproducibility Notes

Implementation. DDO's gradient optimization through the defended model is implemented with `nnsight` [26], which provides differentiable access to model internals while keeping base-model weights frozen. Hyperparameter search uses Optuna [25] with the default Tree-structured Parzen Estimator (TPE) sampler.

We note three reproducibility details. (i) **Determinism:** we use deterministic decoding (`do_sample=false`) and fixed RNG seeds for all stochastic components, and adaptive multi-phase RFA is deterministic (verified across 3 runs). (ii) **Seed sensitivity:** DIM direction estimation introduces $\approx \pm 1\%$ ASR variation across probe-set seeds, and Heretic results vary with Optuna random seed ($\approx \pm 3\%$ across 3 independent 200-trial runs). (iii) **Fixed seeds:** data splits $= 42$, v_proj conditioning $= 99$ (searched over $\{42, 77, 99, 123, 200\}$), KV strengthening $= 7$, Optuna TPE $= 42$.

F.14 Defense and Attack Surface Map

Figure 5 visualizes a single pre-norm transformer block (Llama-3) with defense and attack annotations. DDO edits $\mathbf{W}_{\text{gate}}$, $\mathbf{W}_{\text{up}}$, and $\mathbf{W}_{\text{down}}$ in the SwiGLU MLP (green); RFA hooks activations at three points (red dots); Heretic edits $\mathbf{W}_o$ and $\mathbf{W}_{\text{down}}$ offline (red outlines). The partial overlap on $\mathbf{W}_{\text{down}}$ is the only shared surface between DDO and Heretic.

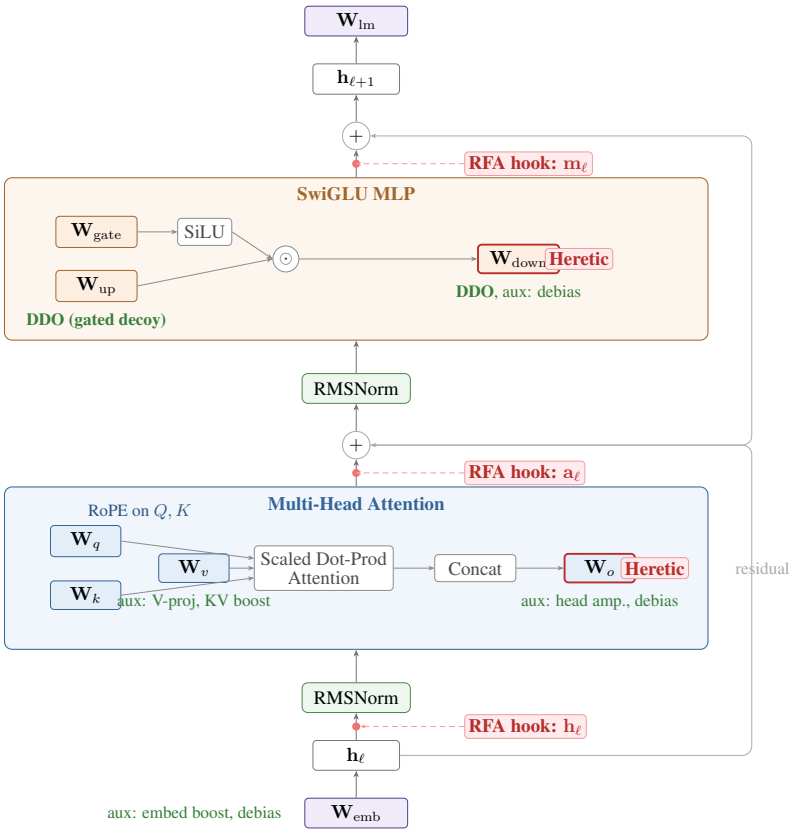

Figure 5: Pre-norm transformer block (Llama-3) annotated with defense and attacker surfaces. Green: **DDO** edits $\mathbf{W}_{\text{gate}}$, $\mathbf{W}_{\text{up}}$, and $\mathbf{W}_{\text{down}}$ in the SwiGLU MLP; auxiliary ("aux") mechanisms optionally edit additional matrices in attention, embeddings, and normalization (Section 3). Red: attacker access—RFA hooks activations at three points ($\mathbf{h}_\ell$, $\mathbf{a}_\ell$, $\mathbf{m}_\ell$); Heretic [14] edits only $\mathbf{W}_o$ and $\mathbf{W}_{\text{down}}$. Overlap occurs only on $\mathbf{W}_{\text{down}}$.

F.15 Additional Model Evaluation

Beyond the six primary models evaluated in the main text (Table 2 and Table 3), we apply DDO with per-model Optuna tuning ($\sim$10 min each) to seven additional instruction-tuned checkpoints spanning 7B–24B parameters (Table 20), including Llama-2-7B-Chat. DDO achieves $<$10% standard-RFA ASR on all seven while preserving MMLU within 1–2 points of the base model, confirming cross-family and cross-scale generalization.

F.16 DDO Undo Attack Analysis

An attacker who obtains only the defended checkpoint (without the original base model) might attempt to identify and revert DDO-modified neurons. A natural undo heuristic is to score each MLP gate row by its projection onto the attacker's estimated refusal direction, $s_i \triangleq \left|\mathbf{w}_{\text{gate}}^{(i)} \cdot \hat{\mathbf{r}}_{\text{atk}}\right|$, where $\hat{\mathbf{r}}_{\text{atk}}$ is the DIM direction estimated on the defended checkpoint, and then zero the top-scoring neurons. DDO's gate rows are shaped to align with the *true* refusal direction $\hat{\mathbf{r}}_{\text{clean}}$ (so the trigger fires on harmful inputs), which would make them easy to find—if $\hat{\mathbf{r}}_{\text{atk}}$ matched $\hat{\mathbf{r}}_{\text{clean}}$. Estimator corruption rotates $\hat{\mathbf{r}}_{\text{atk}}$ away from $\hat{\mathbf{r}}_{\text{clean}}$, frustrating this undo strategy.

Table 20: DDO on additional models with per-model Optuna tuning ($\sim$10 min, single A100 GPU). MMLU: base $\rightarrow$ DDO-defended (%↑). RFA ASR = mean of 4 variants, LlamaGuard-2 (%↓). XST = XSTest (%↑). DDO achieves $<$10% ASR on all models with $\leq$1.4 point MMLU drop.

Model	Size	MMLU↑	RFA↓	XST↑
Llama-2-7B-Chat	7B	48.3→46.9	1.0	79.4
Qwen2.5-7B-Instruct	7B	74.2→73.1	6.0	94.0
Qwen2.5-14B-Instruct	14B	79.7→78.4	8.0	96.0
Qwen3-14B	14B	79.0→77.8	7.0	97.0
Mistral-Small-3-24B	24B	77.5→76.6	2.0	96.0
Mistral-v0.2-7B-Instruct	7B	60.1→59.3	4.0	88.0
Ministral-8B-Instruct	8B	68.0→66.8	5.0	86.0

Mid-layer camouflage. Across Qwen3, Mistral-7B, and Yi, DDO target layers in the mid-zone show $\cos(\hat{\mathbf{r}}_{\text{clean}}, \hat{\mathbf{r}}_{\text{atk}}) \in [-0.05, +0.05]$: the deception is essentially complete. The modified row's projection s_{mod} collapses to within the natural top-5 distribution:

- **Qwen3 L14**: $s_{\text{mod}} = 0.267$ vs. natural rank-2 at 0.230 ($1.16\times$ margin).
- **Mistral-7B L10**: $s_{\text{mod}} = 0.041$, outranked by a natural neuron at 0.066 (rank 4).
- **Yi L26**: $s_{\text{mod}} = 0.112$ vs. natural rank-2 at 0.075 ($1.48\times$ margin).
- **Yi L32**: $s_{\text{mod}} = 0.095$ vs. natural rank-2 at 0.091 ($1.04\times$ margin).

Instruction-tuned models naturally contain MLP neurons whose gate rows align with refusal at the 0.07–0.20 level—these are real refusal-relevant neurons doing work in the model's native safety circuitry. Once $\hat{\mathbf{r}}_{\text{atk}}$ has been rotated away from $\hat{\mathbf{r}}_{\text{clean}}$, DDO's modified neuron sits inside this distribution rather than as an outlier. The undo strategy therefore degrades to "zero the neurons that score highest on the estimated refusal direction"—which is functionally equivalent to standard RFA and provides no advantage over the attacks we already evaluate.

Late-layer visibility. At the deepest target layers, deception can be slightly weaker. Yi L37 (the deepest layer in Yi's L14–37 zone) shows $\cos(\hat{\mathbf{r}}_{\text{clean}}, \hat{\mathbf{r}}_{\text{atk}}) \approx 0.09$, and $s_{\text{mod}} = 0.444$ exceeds the natural rank-2 (0.081) by $\sim 5\times$. However, since the attacker must search the full model without knowing which layers were modified, even a partial margin at one layer does not yield reliable undo across the model. Placing DDO inside the model's natural causal refusal zone gives the defense a stealth property: the modification shares its surface signature with the neurons it is functionally hiding among.

Post-attack ASR under undo heuristics. To complement the geometric analysis above, we measure post-attack ASR after applying each undo heuristic followed by standard RFA. For each heuristic, we zero the top-m down-projection columns per DDO layer, then run 3-point RFA (Table 21).

Table 21: Post-attack ASR (%↓, LlamaGuard-2, 3-point RFA, JailbreakBench) after applying undo heuristics on Llama-3-8B DDO. m = number of neurons zeroed per DDO layer. Base (no defense) ASR = 80%. No heuristic recovers base-level ASR; the best heuristic reaches only 12–17%.

Heuristic	m=1	m=2	m=4	m=8	m=16	m=32
gate·$\hat{\mathbf{r}}$ projection	16	17	15	15	12	13
gate/up cosine	17	14	16	15	12	14
down-col norm	4	7	5	5	4	4

The gate·$\hat{\mathbf{r}}$ and gate/up cosine heuristics partially undo DDO (ASR rises to 12–17%) but remain far below the undefended base (80%), confirming that estimator corruption prevents recovery. The down-column norm heuristic has essentially no effect (4–7% ASR) because DDO selects low-norm neurons whose down-projection columns contribute minimally to model output; zeroing them removes neither the gate/up trigger nor the decoy signal.

Norm-based detection. A detector could flag DDO neurons by their gate/up norm ratio. In the default compile mode, $\mathbf{w}_{\text{gate}} = \beta\hat{\mathbf{r}}$ (norm β) and $\mathbf{w}_{\text{up}} = \hat{\mathbf{r}}$ (norm 1), yielding a ratio of $\beta \approx$ 3–6—an outlier among $\sim$12K neurons. However, this is an artifact of a particular compile

choice, not a fundamental limitation. The trigger output depends on the *product* of gate and up norms: $\mathrm{silu}(\mathbf{w}_{\mathrm{gate}}^{\top}\tilde{\mathbf{x}}) \cdot \mathbf{w}_{\mathrm{up}}^{\top}\tilde{\mathbf{x}}$. Redistributing the total gain—e.g., setting both to $\sqrt{\beta}\,\hat{\mathbf{r}}$—equalizes norms while preserving the trigger in the large-activation regime ($\mathrm{silu}(x) \approx x$): $\mathrm{silu}(\sqrt{\beta}\,a) \cdot \sqrt{\beta}\,a \approx \beta a^2$, matching the original. Similarly, $\mathrm{cossim}(\mathbf{w}_{\mathrm{gate}}, \mathbf{w}_{\mathrm{up}}) = 1.0$ can be reduced by adding a small orthogonal perturbation $\varepsilon\,\mathbf{v}$ to $\mathbf{w}_{\mathrm{up}}$, introducing $O(\varepsilon)$ trigger noise absorbed by the down-projection scale. DDO's free scalar parameters (β, s) provide the degrees of freedom to match any target norm and cosine distribution analytically. All models evaluated in this paper use the default compile without norm redistribution or cosine camouflage.

F.17 Rank-k Attack Sweep

To test robustness beyond rank-1 RFA, we evaluate rank-k SVD attacks: the attacker computes the top-k left singular vectors of the harmful–safe mean activation difference on the defended model and ablates all k directions simultaneously (Table 22).

Table 22: Rank-k attack sweep on Llama-3-8B DDO (ASR %↓, LlamaGuard-2, 3-point RFA, JailbreakBench). DDO protects against rank $\leq$8 attacks; beyond rank 9 (the spectral budget), ASR rises monotonically as the attacker recovers the refusal subspace.

Attack rank k	1	2	4	8	16	32
ASR (%)	4	16	21	29	39	79

DDO provides approximately 9 ranks of spectral budget (consistent with 9 DDO layers $\times$ 1 decoy neuron each). At $k{\leq}8$, the attacker's ablation subspace is dominated by decoy directions and ASR remains low. At $k{=}16$, enough rank is available to substantially recover the refusal subspace (39% ASR). At $k{=}32$, the attack ablates nearly all residual-stream structure (79% ASR), approaching the undefended base (80%). This validates Theorem 2: the attacker must exhaust the decoy rank budget before reaching genuine refusal.

F.18 Probe-Budget Sensitivity

A natural concern is that DDO's estimator corruption might rely on the attacker having too few probes for accurate DIM estimation. We test this by varying the attacker's probe budget from 32 to 1024 harmful+safe pairs (Table 23).

Table 23: Probe-budget sweep on Llama-3-8B DDO (ASR %↓, LlamaGuard-2, 3-point RFA, JailbreakBench). N = number of harmful+safe probe pairs used for DIM estimation. DDO used 128 probes during optimization. ASR is flat across a 32$\times$ range, confirming that DDO's defense is not sample-limited.

Attacker probes N	32	64	128	256	512	1024
ASR (%)	7	6	8	12	7	7

ASR remains flat at 6–8% across a 32$\times$ range of probe budgets (32 to 1024), with a transient bump to 12% at $N{=}256$ that does not sustain. Even with 1024 probes (8$\times$ the DDO training budget of 128), the attacker cannot recover the true refusal direction. This confirms that DDO's defense operates by genuinely corrupting the mean-difference direction, not by exploiting finite-sample noise in the attacker's estimator.

F.19 Licenses for Existing Assets

All models, benchmarks, and software used in this work are publicly available under permissive or research-friendly licenses. We use each asset in accordance with its stated terms.

Models.

- **Llama-3-8B-Instruct** [30][1], **Llama-2-7B-Chat** [61][2], **LlamaGuard-2** [39][3]: Meta Llama 3 Community License (Llama-3, LlamaGuard-2); Llama 2 Community License (Llama-2). Both permit research and commercial use.
- **Gemma-2-9B-IT** [33][4]: Gemma Terms of Use (permissive, allows research and redistribution).
- **Qwen3-8B** [32][5]: Apache 2.0.
- **Mistral-7B-Instruct-v0.3** [34][6]: Apache 2.0.
- **Yi-1.5-9B-Chat** [31][7]: Apache 2.0.
- **GLM-4-9B-Chat** [35][8]: GLM-4 License (permissive for research).
- **GPT-4** (OpenAI): used as MT-Bench judge via API, under OpenAI's Terms of Use.
- **GPT-4o** (OpenAI): used as XSTest and StrongREJECT judge via API, under OpenAI's Terms of Use.

Benchmarks and datasets.

- **JailbreakBench** [24]: MIT License.
- **HarmBench** [5]: MIT License.
- **MMLU** [36]: MIT License.
- **MT-Bench** [37]: Apache 2.0 (part of FastChat / lm-sys).
- **XSTest** [38]: CC-BY-4.0.
- **StrongREJECT** [40]: MIT License.
- **AdvBench** [21]: MIT License.

Software and tools.

- **Heretic** [14]: MIT License.
- **LightEval**: Apache 2.0 (Hugging Face).
- **Optuna**: MIT License.
- **PyTorch**: BSD-style license.
- **Transformers (Hugging Face)**: Apache 2.0.

[1] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[2] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[3] https://huggingface.co/meta-llama/Llama-Guard-2-8B
[4] https://huggingface.co/google/gemma-2-9b-it
[5] https://huggingface.co/Qwen/Qwen3-8B
[6] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[7] https://huggingface.co/01-ai/Yi-1.5-9B-Chat
[8] https://huggingface.co/THUDM/glm-4-9b-chat

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions: (i) DDO as a post-hoc defense, (ii) estimator-corruption theorem and spectral toll, and (iii) comparative evaluation across six model families. All claims are supported by experimental results (Section 4.1) and theoretical analysis (Section 3). Limitations are discussed in Section 5 and Appendix A.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5 and in detail in Appendix A.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theorem 1 (estimator corruption) and Theorem 2 (decoy spectral toll) state all assumptions explicitly. Corollary 3 (flat spectrum) is in the main text; Corollary 4 (decoy-dominated contrast) is in the appendix. Complete proofs are in Appendix E.1.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 specifies models, inference settings, coherence checks, data splits, benchmarks, and judges. Appendix F.2 provides exact hyperparameters for all mechanisms. Appendix F.1 details attack implementations. Appendix F.6 provides per-model DDO configurations. Reproducibility details (deterministic decoding, fixed seeds, verified determinism of adaptive RFA) are documented in the appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Anonymized code is available at `https://anonymous.4open.science/r/mechanistic-defenses/`. All models used are publicly available open-weight checkpoints. Evaluation benchmarks (JailbreakBench, HarmBench, MMLU, MT-Bench, XSTest) are publicly available. Detailed hyperparameters and implementation details are provided in the appendix to enable independent reimplementation.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes data splits (128 harmful + 128 safe probes for estimation, held-out test sets for evaluation), inference settings (deterministic decoding, batch size 8, max 512 tokens), and baseline configurations. Appendix F.2 provides full hyperparameters. Appendix F.6 reports per-model DDO configurations with Optuna search spaces. Appendix F.8 details baseline sweep procedures.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We use deterministic decoding and fixed seeds, so individual runs are fully reproducible. We report three forms of statistical information: (i) ASR is averaged over three independent judges (HarmBench classifier, LlamaGuard-2, StrongREJECT) to reduce single-judge variance; (ii) adaptive multi-phase RFA is verified deterministic across 3 identical runs; and (iii) the appendix reports that DIM direction estimation introduces $\approx \pm 1\%$ ASR variation across probe-set seeds.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 4.1 states all experiments use a single A100 GPU. The appendix (Compute Budget table) provides wall-clock times for each step: DDO tuning ($\sim$5 min), full mechanism optimization ($\sim$25 min), RFA evaluation ($\sim$5 min), Heretic evaluation ($\sim$30–60 min), and MT-Bench per phase ($\sim$20 min). The cross-model baseline comparison required training 4 defenses $\times$ 3–4 configs $\times$ 5 models ($\sim$15 A100-hours each).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This work develops defenses against attacks on LLM safety mechanisms. We do not release uncensored models or tools for removing safety guardrails. All harmful prompt datasets used are from established public benchmarks (JailbreakBench, HarmBench, AdvBench). The research aims to improve the safety of open-weight LLM releases.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Positive and negative societal impacts, dual-use concerns, and responsible disclosure are discussed in Appendix A.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release new models, datasets, or tools for attacking LLM safety. DDO is a defensive tool that hardens existing open-weight checkpoints. All models used are publicly available, and all evaluation benchmarks are established public resources.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models, benchmarks, and tools are cited with their original publications, and their licenses are listed in Appendix F.19. Baseline defense checkpoints are cited and sourced (Section 4.1). Heretic is cited with its open-source repository. All evaluation frameworks (HarmBench, JailbreakBench, MT-Bench, MMLU, XSTest, nnsight) are cited.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new datasets or pre-trained models. DDO code is available at the anonymized repository linked above.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing or human subjects research. All evaluations are automated using established benchmarks and LLM-based judges.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve human subjects research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are central to this work as the objects of study: we evaluate defense and attack methods on instruction-tuned LLMs (Llama-3, Gemma-2, Qwen3, etc.). GPT-4 is used as an MT-Bench judge, GPT-4o as a StrongREJECT rubric evaluator, and LlamaGuard-2 as an ASR judge. These uses are documented in Section 4.1. Additionally, Claude Opus 4.6 was used to propose hyperparameters for trained baselines (Config E; Appendix F.8).