DEFENSES AT ODDS: MEASURING AND EXPLAINING DEFENSE CONFLICTS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) may require additional defenses after deployment as risks and governance requirements evolve. Subsequent defenses can interact with earlier defenses, raising the question of their sequential compatibility. We study this question with CONFLICTEVAL, evaluating 144 ordered compositions of six defenses spanning safety, privacy, and fairness across six models. The resulting interactions are heterogeneous across defense pairs, application orders, and models. Notably, some compositions exhibit defense conflicts: the subsequent defense improves its target objective while weakening protection established by the earlier defense. We investigate these interactions through the directional compatibility of defense-induced changes in risk-relevant representations. Across 11 selected cases, we use activation interventions to assess which defense-induced representational changes support protection and examine how subsequent defenses affect these changes. We find that, in some conflict cases, subsequent defenses counteract representational changes supporting earlier protection, providing evidence for one possible pathway to defense conflicts. Building on this analysis, we propose *Conflict-Triggered Directional Retention* (CTDR), which penalizes opposing shifts along directions supporting earlier protection. On six selected conflicting compositions from this analysis, CTDR reduces first-objective regression while maintaining positive gains on the subsequent defense objective.

1 INTRODUCTION

Post-training defenses help large language models (LLMs) meet safety, privacy, and fairness requirements. Approaches such as safety alignment (Rafailov et al., 2023; Ouyang et al., 2022; Xhonneux et al., 2024), machine unlearning (Bourtoule et al., 2021; Zhang et al., 2024; Li et al., 2024b), and debiasing (Ilharco et al., 2022) aim to reduce harmful responses, limit memorized-data leakage, and mitigate stereotypical biases, respectively. However, LLMs may require additional defenses after deployment as new vulnerabilities emerge (Google DeepMind, 2024; OpenAI, 2025). For example, a model already aligned for safety may later receive a data-removal request, requiring an update that reduces memorized information while preserving existing protections. Such evolving requirements can lead to defenses being applied sequentially, with each update acting on a model already shaped by earlier defenses. This raises a practical question: ***when a subsequent defense improves its target objective, does it preserve, strengthen, or weaken the protection established by earlier defenses?***

A defense's effectiveness on its target risk alone does not establish its compatibility with protections already in place. Prior work (Qi et al., 2023; Bhardwaj et al., 2024; Qi et al., 2025) has shown that subsequent fine-tuning, even on benign tasks, can weaken the safety alignment of LLMs. When the subsequent update is itself a defense, it may improve its target objective while weakening protection established by an earlier defense. We refer to this outcome as a *defense conflict*. Such conflicts highlight the need to evaluate the gains from a subsequent defense alongside the retention of earlier protection.

To examine sequential defense compatibility, we introduce CONFLICTEVAL and evaluate 144 ordered compositions of six defenses spanning safety, privacy, and fairness across six models. We jointly evaluate the subsequent defense's target-objective gain and the Relative Regression Rate (RRR), which quantifies how much of the first defense's achieved risk reduction is reversed by the subsequent update. Our evaluation reveals heterogeneous interactions across defense pairs, applica-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion orders, and models: subsequent defenses can strengthen earlier protection in some configurations while weakening it in others. Notably, some compositions exhibit defense conflicts in which the subsequent defense improves its own target objective while eroding earlier protection, showing that an effective defense update can come at the expense of protection already established.

We investigate a possible representational explanation for these conflicts: subsequent defenses may weaken earlier protection by counteracting defense-induced changes in risk-relevant representations. Across 11 selected cases, we use activation interventions (Meng et al., 2022) to assess which defense-induced changes along risk-relevant directions contribute to protection, and examine how subsequent updates affect them. In some conflict cases, later updates oppose these defense-supporting shifts, providing evidence for representational counteraction as one pathway to the observed protection loss. Building on this analysis, we propose *Conflict-Triggered Directional Retention* (CTDR), which penalizes opposing representational shifts along directions supporting earlier protection during subsequent defense training. We evaluate CTDR on six selected conflicting compositions from the mechanism analysis, where it reduces first-objective regression while maintaining positive gains on the subsequent defense objective.

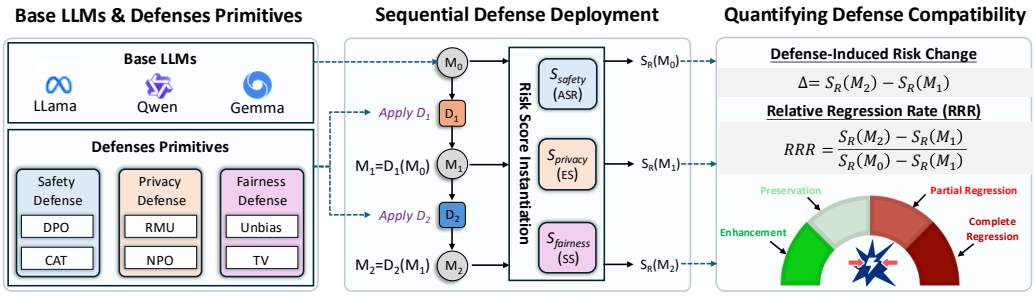


Figure 1: Overview of the CONFLICTEVAL evaluation framework, which quantifies cross-defense interference under sequential deployment ($D_1 \rightarrow D_2$) across safety, privacy, and fairness dimensions via the Relative Regression Rate (RRR).

2 MEASURING SEQUENTIAL DEFENSE COMPATIBILITY

2.1 EVALUATION FRAMEWORK

To examine sequential defense compatibility, we introduce CONFLICTEVAL, an evaluation framework that jointly measures the gains from a subsequent defense and the retention of protection established by an earlier defense. We instantiate the framework with six defenses spanning safety, privacy, and fairness across six models, yielding 144 ordered compositions. As illustrated in Figure 1, the framework tracks both target risks across successive model checkpoints and quantifies whether new defense gains accompany the preservation or regression of earlier protection.

Sequential composition. We consider two configured defense procedures, D_1 and D_2, applied sequentially. In this study, we focus on cross-risk compositions in which the two defenses target distinct risk dimensions, $R_1 \neq R_2$. Starting from an initial model M_0, the first defense produces M_1, and the subsequent defense updates this already defended model to obtain M_2:

$$M_1 = D_1(M_0), \qquad M_2 = D_2(M_1). \tag{1}$$

Here, each defense denotes its configured algorithm, training data, hyperparameters, and update procedure. Thus, an ordered composition $D_1 \rightarrow D_2$ specifies both the two instantiated defenses and the sequence in which they are applied to the model.

Quantifying compatibility. Let $S_R(M)$ denote the risk score of model M on dimension R, with larger values indicating higher risk. For each risk dimension, all three checkpoints are evaluated on the same fixed evaluation set. To respect this sequential setup, we measure the gain of each defense relative to the checkpoint from which it is applied:

$$\begin{aligned} G_1 &= S_{R_1}(M_0) - S_{R_1}(M_1), \\ G_2 &= S_{R_2}(M_1) - S_{R_2}(M_2). \end{aligned} \tag{2}$$

Positive G_i indicates a reduction in the corresponding target risk. In particular, G_2 is measured from M_1 rather than M_0, so that it captures the incremental change associated with D_2 rather than changes already induced by D_1. The subsequent defense's effect on earlier protection is

$$\Delta_1 = S_{R_1}(M_2) - S_{R_1}(M_1), \tag{3}$$

where positive Δ_1 indicates regression of earlier protection and negative Δ_1 indicates further risk reduction. Because the same increase in R_1 can reverse different fractions of the risk reduction achieved by D_1, we normalize Δ_1 by G_1 to obtain the Relative Regression Rate:

$$\mathrm{RRR} = \frac{\Delta_1}{G_1}. \tag{4}$$

RRR is defined only when $G_1 > \tau$, where $\tau = 10^{-8}$ is a numerical tolerance. Positive RRR indicates regression of the first-defense gain, with larger values corresponding to greater regression; $\mathrm{RRR} = 1$ corresponds to a return to the initial risk level. Negative RRR indicates further reduction of the earlier risk. We identify an observed *defense conflict* when the subsequent defense reduces its own target risk while protection established by the earlier defense regresses.

2.2 Evaluation Setup

Models and initialization. We evaluate six instruction-tuned models from the Llama, Gemma, and Qwen families, selecting a smaller (S) and a larger (L) model from each family (see Table 1). To provide a controlled setting for evaluating privacy defenses, we adopt the TOFU initialization procedure used in unlearning benchmarks (Maini et al., 2024; Dorna et al., 2025). TOFU contains question–answer pairs about fictitious authors, allowing models to acquire identifiable target knowledge through fine-tuning before unlearning is applied. We fine-tune each model on this dataset and use the resulting checkpoint as M_0. All defense compositions for a given model start from this same checkpoint, enabling comparisons of sequential updates from a common initial state.

Table 1: Target LLMs used in ConflictEval. S and L denote the smaller and larger models.

Model family	Short name	Model identifier
Llama	Llama-S	Llama-3.2-1B-Instruct (Meta Llama Team, 2024a)
	Llama-L	Meta-Llama-3-8B-Instruct (Meta Llama Team, 2024b)
Gemma	Gemma-S	gemma-2-2b-it (Google, 2024a)
	Gemma-L	gemma-7b-it (Google, 2024b)
Qwen	Qwen-S	Qwen2.5-1.5B-Instruct (Qwen Team, Alibaba Group, 2024a)
	Qwen-L	Qwen2.5-7B-Instruct (Qwen Team, Alibaba Group, 2024b)

Defenses and coverage. We instantiate two defenses for each risk dimension, chosen to cover distinct update mechanisms. The safety defenses are DPO (Rafailov et al., 2023), based on preference optimization over safe and unsafe responses, and CAT (Xhonneux et al., 2024), based on adversarial training with continuous input perturbations. The privacy defenses are NPO (Zhang et al., 2024), which applies a negative preference objective to forget data, and RMU (Li et al., 2024b), which modifies representations associated with targeted content. For fairness, we use Unbias (Liu et al., 2025), which applies targeted unlearning to stereotypical associations, and task-vector editing (TV (Ilharco et al., 2022; Xu et al., 2025)), which counteracts a bias-related parameter update. For each model, we evaluate every ordered pair of defenses targeting different risk dimensions, yielding $6 \times 4 = 24$ ordered compositions per model and 144 overall. Unless otherwise stated, sequential compositions apply the configured defense procedures without any additional mechanism designed to preserve the protection established by D_1. See Appendix A for defense implementation details.

Risk Score Instantiation. We instantiate S_R separately for safety, privacy, and fairness using benchmark-specific metrics. All risk scores are oriented so that larger values indicate higher risk.

- **Safety Risk.** We measure safety risk using the *Attack Success Rate* (ASR) on 404 high-risk prompts sourced from SALAD-Bench (Li et al., 2024a). The prompts were drawn from the Malicious Use and Representation & Toxicity categories (excluding Unfair Representation) and retained if they elicited a harmful response from at least one of the six M_0 checkpoints. A response

Table 2: Relative Regression Rate (RRR) across all ordered defense compositions. For positive RRR, darker red indicates greater regression of the first-defense gain; color intensity increases linearly from $\mathrm{RRR} = 0$ to $\mathrm{RRR} = 1$. Color intensity increases linearly from RRR = 0 to RRR = 1. Non-positive values are left unshaded.

Deployment Scenario	Primary Defense (D_1)	Subsequent Defense (D_2)	Llama		Gemma		Qwen	
			-S	-L	-S	-L	-S	-L
Fairness-First	**Unbias** (Fairness Stage)	+ RMU (Privacy)	−0.005	−0.152	0.154	0.215	0.027	0.072
		+ NPO (Privacy)	−0.401	−0.222	0.174	−0.316	0.012	0.201
		+ DPO (Safety)	−0.284	−0.181	0.403	−0.300	0.048	−0.078
		+ CAT (Safety)	0.600	0.724	0.637	0.301	0.659	0.526
	Task Vector (TV) (Fairness Stage)	+ RMU (Privacy)	0.014	−0.134	−0.887	0.389	−0.091	−0.495
		+ NPO (Privacy)	0.011	−0.169	0.311	−0.0006	0.029	0.058
		+ DPO (Safety)	0.091	−0.001	−0.438	0.158	0.018	−0.011
		+ CAT (Safety)	0.403	0.346	0.575	0.188	0.609	0.218
Privacy-First	**NPO** (Privacy Stage)	+ Unbias (Fairness)	−0.002	−0.564	0	0	−0.078	−0.0006
		+ TV (Fairness)	0.125	−0.639	0	0	−0.079	−0.001
		+ DPO (Safety)	−0.003	−0.463	0	0	−0.078	0.0004
		+ CAT (Safety)	0.045	−0.656	0.037	0.016	−0.009	0.045
	RMU (Privacy Stage)	+ Unbias (Fairness)	−0.059	−0.277	−0.170	−0.190	−0.379	−0.104
		+ TV (Fairness)	−0.057	−0.255	−0.171	−0.196	−0.869	−0.224
		+ DPO (Safety)	−0.058	−0.261	0.022	−0.158	0.017	−0.070
		+ CAT (Safety)	0.053	−0.250	−0.055	−0.181	−0.356	−0.130
Safety-First	**DPO** (Safety Stage)	+ Unbias (Fairness)	−0.265	−0.968	0.798	0.053	−0.146	−0.974
		+ TV (Fairness)	−0.473	−1.879	0.505	−0.391	−1.457	−2.864
		+ NPO (Privacy)	0.228	−0.681	0.697	−0.314	−0.477	−0.064
		+ RMU (Privacy)	0.310	−0.315	0.132	0.008	−0.132	−0.206
	CAT (Safety Stage)	+ Unbias (Fairness)	−0.120	−0.105	−0.361	0.409	0.039	0.002
		+ TV (Fairness)	−0.094	0.032	−0.264	−0.348	0.301	0.417
		+ NPO (Privacy)	−0.101	−0.187	0.090	0.222	0.085	−0.004
		+ RMU (Privacy)	−0.173	−0.011	−0.097	0.038	−0.069	−0.003

is counted as harmful when its MD-Judge v0.2 [1] score exceeds 3. We average ASR over three decoding seeds using temperature 0.6, top-p 0.9, and a maximum of 256 generated tokens.

- **Privacy Risk.** We instantiate privacy risk using mean extraction strength (ES) on the TOFU forget set (Maini et al., 2024), following OpenUnlearning's teacher-forced exact-suffix matching evaluator (Dorna et al., 2025). ES measures matching answer suffixes between tokenwise predictions and the target answer, with higher values indicating greater residual target-content predictability.
- **Fairness Risk.** We measure fairness risk using StereoSet intrasentence examples (Nadeem et al., 2021). Let $\mathrm{SS}(M)$ denote the stereotype preference score, expressed as a percentage. We define $S_{\mathrm{fairness}}(M) = |\mathrm{SS}(M) - 50|/50$, the normalized deviation from the balanced point $\mathrm{SS} = 50$, such that larger values indicate stronger preference in either direction.

These benchmark-specific risk scores form the basis for subsequent compatibility measurements. Defense gains and RRR values are computed from unrounded checkpoint scores.

2.3 Observed Sequential Compatibility

Table 2 summarizes the observed compatibility of the 144 ordered compositions. Full risk scores at each checkpoint are reported in Appendix B. We additionally evaluate MMLU (Hendrycks et al., 2021) as a utility check; most compositions remain close to their undefended baselines, suggesting that the observed risk regressions are not attributable to broad capability collapse (Appendix C).

Defense conflicts are observed across the evaluated compositions. Across the 144 ordered compositions, 57 meet our definition of a defense conflict: the subsequent defense reduces its target risk

[1] https://huggingface.co/OpenSafetyLab/MD-Judge-v0_2-internlm2_7b

while the risk addressed by the earlier defense increases. Observed conflicts span all three deployment scenarios, occurring in 30 of 48 fairness-first, 18 of 48 safety-first, and 9 of 48 privacy-first compositions. Because each configured composition is represented by a single training realization, these counts are descriptive of the observed runs rather than estimates of how frequently the same configurations would exhibit conflicts across random seeds. We therefore use them to characterize the breadth of observed defense conflicts, not their seed-level prevalence.

Observed compatibility varies across defense pairs and models. The observed compatibility patterns are not determined solely by the pair of risk dimensions being defended. For example, on Llama-S, applying NPO after DPO increases safety risk from 0.4670 to 0.5095 ($\mathrm{RRR} = 0.228$), whereas applying NPO after CAT decreases risk from 0.2954 to 0.2591 ($\mathrm{RRR} = -0.101$). In both cases, NPO reduces privacy risk, but its effect on the protection established by the preceding safety defense differs. The same ordered composition can also produce different outcomes across models. When NPO is applied after DPO, safety protection regresses on Gemma-S ($\mathrm{RRR} = 0.697$), whereas on Gemma-L the same composition further reduces safety risk from 0.2673 to 0.1444 ($\mathrm{RRR} = -0.314$), while reducing privacy risk in both cases. These observations show that sequential compatibility cannot be inferred from the targeted risk dimensions alone; within our evaluation, it varies across defense pairs and models.

Observed compatibility varies with application order. The same two defenses can exhibit different compatibility when their order is reversed. On Llama-S, applying RMU after DPO causes regression of the preceding safety protection ($\mathrm{RRR} = 0.310$), whereas applying DPO after RMU further reduces the privacy risk established by RMU ($\mathrm{RRR} = -0.058$). Thus, compatibility need not be symmetric: the behavior of one ordering does not determine the reverse. This motivates evaluating defense compositions as ordered sequences rather than unordered pairs.

Taken together, these measurements show that defense conflicts can arise under sequential defense deployment and interactions are heterogeneous across defense pairs, application orders, and models.

3 Representational Analysis of Defense Conflicts

The behavioral results motivate a representational hypothesis: subsequent defenses may weaken earlier protection by counteracting representational changes that support it. We investigate this hypothesis by identifying risk-relevant representations, assessing the contribution of defense-induced changes along these representations, and measuring how subsequent updates affect those changes.

3.1 Analysis Setup

Case selection. We analyze 11 selected cases from the sequential-composition evaluation (Table 3). These cases include compositions in which subsequent defenses weaken, preserve, or strengthen earlier protection, allowing us to examine representational changes across these different outcomes. They span three model families, both model scales, and all three risk dimensions. Two reverse-order pairs allow us to compare how the same defenses interact under different application orders.

Analysis data. We use examples from the evaluation datasets described in Section 2.2 to construct the risk-specific contrasts defined below. We partition the analysis data into discovery, development, and test sets: discovery data determine candidate risk-relevant representations and intervention scales, development data select directions and defense-support locations, and test data evaluate these fixed selections and the representational changes under sequential composition.

3.2 Identifying Risk-Relevant Representations

We characterize risk-relevant representations as directions in activation spaces. To make intervention evaluation computationally efficient, we use risk proxy scores based on risk-token log probabilities, which capture model preferences for risk-relevant text without full response generation.

Estimating candidate risk directions. Let $h_\ell(M, x)$ denote the residual activation of input x at Transformer block ℓ. For safety, we use the activation at the final prompt token. For privacy and

fairness, we average activations over the selected predictive positions following divergence between paired candidates.

For each risk dimension R, we construct two discovery-set contrast groups, $\mathcal{D}_R^A$ and $\mathcal{D}_R^B$. The contrasts are harmful versus harmless requests for safety, original versus question-matched perturbed answers for privacy, and stereotypical versus anti-stereotypical sentences for fairness. We estimate the candidate direction as

$$d_{R,\ell} = \frac{1}{|\mathcal{D}_R^A|} \sum_{x \in \mathcal{D}_R^A} h_\ell(M_0, x) - \frac{1}{|\mathcal{D}_R^B|} \sum_{x \in \mathcal{D}_R^B} h_\ell(M_0, x). \tag{5}$$

We normalize $d_{R,\ell}$ to obtain a unit vector $u_{R,\ell}$, which serves as the risk direction at layer ℓ.

Evaluating candidate risk directions. We assess whether candidate directions influence risk-related model behavior by increasing or decreasing the activation component along each direction. For each $u_{R,\ell}$, we intervene on M_0 at the analysis positions used for direction estimation:

$$h_\ell \leftarrow h_\ell \pm \alpha \sigma_{R,\ell} u_{R,\ell}, \tag{6}$$

where $\alpha = 0.25$ and $\sigma_{R,\ell}$ is the standard deviation across discovery examples of the target-sequence mean projections onto $u_{R,\ell}$. Let $M_0^{+,\ell}$ and $M_0^{-,\ell}$ denote the models under the positive and negative interventions, respectively. We quantify the directional effect as

$$E_{R,\ell} = \bar{q}_R(M_0^{+,\ell}) - \bar{q}_R(M_0^{-,\ell}), \tag{7}$$

where q_R is a risk proxy score derived from token log probabilities. We use this proxy to reduce the cost of repeated intervention evaluations and capture fine-grained changes in model preferences. For safety risk, it measures the difference in mean token log probability between compliance and refusal prefixes. For privacy risk, it measures the mean token log probability of the original answer. For fairness risk, it measures the stereotype–anti-stereotype log-probability margin. We select directions with resolved orientation and positive effects above numerical tolerance on development data.

3.3 Assessing Defense-Supporting Changes

We next assess whether defense-induced changes along the identified directions contribute to protection. We test this by restoring their pre-defense projections and measuring whether defense performance deteriorates.

Testing defense support. For defense D_i targeting R_i, we examine its update $M_{i-1} \to M_i$ using the risk direction $u_{i,\ell} = u_{R_i,\ell}$ identified on M_0. At each layer with a direction selected in Section 3.2, we restore the activation projection along that direction to its pre-defense value:

$$h_\ell(M_i, x) \leftarrow h_\ell(M_i, x) - u_{i,\ell} u_{i,\ell}^\top \big[h_\ell(M_i, x) - h_\ell(M_{i-1}, x)\big]. \tag{8}$$

This intervention removes the projected change while preserving orthogonal components at the intervention site. Let $M_i^{\text{restore},\ell}$ denote the model under this intervention. We measure its effect as

$$A_{i,\ell} = \bar{q}_{R_i}(M_i^{\text{restore},\ell}) - \bar{q}_{R_i}(M_i). \tag{9}$$

A positive $A_{i,\ell}$ indicates that restoration increases the risk proxy score, suggesting that the defense relies in part on changes along this risk representation.

Results. We record the layers containing risk-relevant representational changes that support defense D_i on development data as K_i. As shown in Table 3, K_1 and K_2 overlap in 10 of the 11 compositions, sharing between one and eleven layers. Qwen-L Unbias→CAT is the only case with disjoint support sets. Shared layers place the defense-supporting changes in the same activation space, although their directions may differ. We next examine how subsequent updates affect the changes supporting earlier protection.

3.4 Relating Sequential Changes to Regression

We next examine whether subsequent updates reinforce or counteract representational changes supporting earlier protection. We measure their directional relationship along the first-risk direction and compare it with behavioral regression.

Table 3: Defense-support locations and sequential effects in 11 compositions. K_1 and K_2 are selected on development data for $D_1(M_0)$ and $D_2(M_1)$, respectively; entries are one-based block indices. RDC is pooled over $K_1 \cap K_2$, except †, which uses K_1. Negative RDC indicates opposition to the first-defense shift; positive RRR indicates earlier-protection regression.

		Defense-support locations			Sequential effects	
Model	Defense sequence	K_1	K_2	$K_1 \cap K_2$	Pooled RDC	RRR
Llama-S	DPO→RMU	8–10, 12–15	2, 11–13	12, 13	−1.687	0.310
	RMU→DPO	2, 11–13	8–10, 12–15	12, 13	0.389	−0.058
	DPO→NPO	8–10, 12–15	4, 8, 11–13	8, 12, 13	−1.789	0.228
Gemma-S	DPO→NPO	1, 2, 7, 15	11, 15, 17–26	15	−0.050	0.697
	NPO→DPO	11, 15, 17–26	1, 2, 7, 15	15	0.169	0.000
Gemma-L	CAT→NPO	10, 11, 13, 21–28	16, 23	23	−0.099	0.222
	CAT→TV	10, 11, 13, 21–28	5, 12, 24	24	1.206	−0.348
Qwen-S	Unbias→DPO	1, 2, 4, 6, 7, 14, 17, 19–27	4, 23–28	4, 23–27	−0.054	0.048
	TV→NPO	1–3, 5, 6, 14–21, 23, 27	2, 9, 13–27	2, 14–21, 23, 27	−0.014	0.029
Qwen-L	Unbias→CAT	5–10	2, 3, 12, 13, 16, 20, 24, 25, 28	∅	−2.207†	0.526
	NPO→CAT	11, 12, 15, 16, 21, 25	2, 3, 12, 13, 16, 20, 24, 25, 28	12, 16, 25	−1.322	0.045

Measuring directional compatibility. For each input x, we project both sequential updates onto the fixed first-risk direction $u_{1,\ell}$:

$$a_\ell(x) = u_{1,\ell}^\top \big[h_\ell(M_1, x) - h_\ell(M_0, x)\big], \tag{10}$$

$$b_\ell(x) = u_{1,\ell}^\top \big[h_\ell(M_2, x) - h_\ell(M_1, x)\big]. \tag{11}$$

Here, $a_\ell(x)$ captures the first-defense change, while $b_\ell(x)$ captures the subsequent change along the same direction. Their product is negative when the two updates oppose each other. We compute these projections on test examples for the first risk, using the same activation positions as in the preceding analysis.

Let $K = K_1 \cap K_2$ denote the shared support layers and $\mathcal{T}_{R_1}$ the first-risk test set. We define pooled Representational Directional Compatibility (RDC) as

$$\mathrm{RDC}(K) = \frac{\sum_{\ell \in K} \sum_{x \in \mathcal{T}_{R_1}} a_\ell(x) b_\ell(x)}{\sum_{\ell \in K} \sum_{x \in \mathcal{T}_{R_1}} a_\ell(x)^2}. \tag{12}$$

The denominator normalizes the interaction by the squared magnitude of the first-defense projections. Negative RDC indicates aggregate opposition to the first-defense changes, whereas positive RDC indicates alignment. RDC is an uncentered regression coefficient and can exceed one in magnitude. Its magnitude does not directly quantify behavioral regression. For Qwen-L Unbias→CAT, the shared set is empty. We therefore compute RDC over K_1 to examine how CAT affects the changes supporting Unbias, and mark this result separately in Table 3.

Results. Across the selected compositions, representational counteraction closely tracks behavioral regression. All eight cases in which the earlier protection regresses have negative RDC, meaning that the subsequent defense moves the model against the representational changes induced by the first defense. By contrast, the three cases without regression have positive RDC, indicating that the second update largely preserves or reinforces those earlier changes. This pattern is particularly clear in the reverse-order pairs. For Llama-S, RMU counteracts the changes established by DPO and the earlier protection regresses; reversing the order instead yields aligned changes with no regression. Gemma-S shows the same qualitative pattern for DPO and NPO: the order with opposing changes exhibits regression, whereas the reverse order shows aligned changes and no regression. These cases suggest that the interaction depends not simply on which defenses are combined, but on whether the later update reinforces or counteracts the changes established by the earlier defense.

Qwen-L Unbias→CAT illustrates a possible interaction outside shared support layers. Although the two support sets are disjoint, CAT produces negative RDC (-2.207) at the Unbias support layers, alongside behavioral regression. A defense's support set identifies changes contributing to its own

Table 4: CTDR results on six defense conflicts. Sequential denotes composition without CTDR. Retained gives the percentage of sequential G_2 preserved by CTDR. Lower RRR indicates better retention of earlier protection; higher G_2 indicates greater subsequent-defense gain. Red cells mark reference conflicts, and gray cells show CTDR results. † No shared support layers ($K_1 \cap K_2 = \varnothing$).

Model	Defense sequence	RRR ↓		G_2 ↑		
		Sequential	CTDR	Sequential	CTDR	Retained
Llama-S	DPO→RMU	0.310	−0.053	0.636	0.633	99.5%
	DPO→NPO	0.228	−0.434	0.645	0.631	97.8%
Gemma-S	DPO→NPO	0.697	0.000	0.925	0.896	96.9%
Gemma-L	CAT→NPO	0.222	−0.439	0.112	0.112	100.0%
Qwen-L	NPO→CAT	0.045	0.000	0.638	0.641	100.5%
	Unbias→CAT†	0.526	−0.077	0.687	0.693	100.9%

effect, but does not delimit all representations it may modify. This case therefore suggests that a subsequent defense may disrupt earlier protection through changes outside its own identified support layers. Together, these observations support representational counteraction as one possible pathway to defense conflicts and motivate preserving first-defense changes during subsequent training.

4 Mitigating Defense Conflicts

The preceding analysis suggests that subsequent defenses may weaken earlier protection by counteracting changes that support it. We therefore test whether preserving these changes can reduce regression while retaining the benefits of the subsequent defense. We propose *Conflict-Triggered Directional Retention* (CTDR), which penalizes opposing changes along the identified first-risk directions during subsequent defense training.

Let M_θ be initialized from M_1 and trained on the second defense objective. We use fixed first-risk development examples as anchors. For each anchor x and support layer $\ell \in K_1$, we retain the first-defense projection $a_\ell(x)$ defined in Equation 10 and measure the current change as

$$b_\ell(x;\theta) = u_{1,\ell}^\top \big[h_\ell(M_\theta, x) - h_\ell(M_1, x)\big]. \tag{13}$$

A negative product $a_\ell(x)b_\ell(x;\theta)$ indicates that the current change opposes the first-defense shift. For an anchor minibatch $\mathcal{B}$, we optimize

$$\mathcal{L}(\theta) = \mathcal{L}_{D_2}(\theta) + \lambda \sum_{\ell \in K_1} w_\ell \frac{\sum_{x\in\mathcal{B}} \big[-a_\ell(x)b_\ell(x;\theta)\big]_+^2}{\sum_{x\in\mathcal{B}} a_\ell(x)^2 + \epsilon}, \tag{14}$$

where $[z]_+ = \max(z, 0)$, $\mathcal{L}_{D_2}$ is the subsequent defense's training objective, and λ controls the retention penalty. We set $\epsilon = 10^{-8}$ and weight each layer by its positive development restoration effect:

$$w_\ell = \frac{A_{1,\ell}}{\sum_{j\in K_1} A_{1,j}}. \tag{15}$$

Thus, layers with larger restoration effects receive greater weight. The penalty activates only for opposing changes and does not directly constrain aligned changes or orthogonal activation components. We apply it over K_1, including layers outside the shared support set. Reference projections from M_0 and M_1 are cached.

Evaluation setup. We evaluate CTDR on the five stronger conflicts in the mechanism analysis (RRR 0.222–0.697) and one weaker conflict observed when CAT follows NPO on Qwen-L (RRR 0.045), covering both substantial and mild regression. Following Section 2.2, we compare CTDR with the reference sequential compositions using RRR for earlier-protection retention and G_2 for subsequent-defense gains. We use $\lambda = 1.0$ for all six cases and keep the other defense-training settings specified in Appendix A.

Results. CTDR improves the retention of earlier protection while preserving the benefits of subsequent defenses (Table 4). Across the six cases, RRR becomes non-positive or approaches zero,

while approximately 97% or more of the subsequent-defense gain is retained. The improvement extends across both the stronger conflicts and the weaker conflict observed when CAT follows NPO on Qwen-L. The retained gains indicate that the constraint still allows the subsequent defense to make progress on its target objective. In these evaluated settings, much of the earlier-objective regression can therefore be mitigated with only a small reduction, if any, in subsequent-defense effectiveness. Moreover, CTDR also mitigates regression when CAT follows Unbias on Qwen-L, despite their disjoint support sets. RRR decreases from 0.526 to -0.077, with no decrease in the reported subsequent-defense gain. This regression may involve incidental changes to general-purpose representations: large updates or optimization noise during the second defense could perturb representations that also support earlier protection. The mitigation result is consistent with CTDR limiting such interference by constraining opposing changes at the first-defense support layers.

5 Related Work

Sequential updates and defense retention. Prior work has shown that later model updates can degrade behaviors established by earlier training. Downstream fine-tuning can weaken safety alignment (Qi et al., 2023; Bhardwaj et al., 2024), while machine unlearning can be partially reversed by subsequent fine-tuning, motivating methods that improve post-unlearning robustness (Wang et al., 2025). These studies establish that post-hoc optimization can interfere with earlier protections. We study a distinct setting in which the subsequent update is itself an independently motivated defense targeting a different risk dimension, and systematically evaluate such ordered compositions across safety, privacy, and fairness.

Representational analysis of alignment. Prior interpretability work has identified localized and low-dimensional representations associated with aligned behavior, including safety-critical layers (Li et al., 2024c), safety-related activation directions (Pan et al., 2025), and readable activation changes induced by fine-tuning (Minder et al., 2025). Activation interventions further provide tools for testing whether internal representations causally support model behavior (Meng et al., 2022). Building on these ideas, we study not an individual defense in isolation, but how representational changes induced by a subsequent defense interact with changes supporting earlier protection, and use this analysis to motivate directional retention during sequential defense training.

6 Discussion and Conclusion

ConflictEval studies sequential defense compatibility across safety, privacy, and fairness. Across 144 ordered compositions, we find that subsequent defenses can preserve, strengthen, or weaken earlier protection, and that these interactions vary across defense pairs, application orders, and models. In particular, we observe defense conflicts in which a new defense improves its own objective while causing previously established protection to regress, showing that successful defense updates do not necessarily compose safely. Our representational analysis provides evidence that counteraction of defense-supporting changes is one possible pathway to such conflicts, linking behavioral regression to how successive defenses reshape risk-relevant representations. Building on this analysis, CTDR reduces earlier-objective regression while retaining subsequent-defense gains in six evaluated conflicts. Overall, our results highlight that sequential defenses should be evaluated not only by the gains they introduce, but also by how well they preserve protection already established.

Limitations. Our study has several limitations. First, the 144 compositions are evaluated with a single defense-training realization, so the observed conflicts should not be interpreted as stable properties of particular defense pairs across random seeds. This limitation does not affect our primary goal of establishing that defense conflicts can arise under sequential deployment and that their occurrence depends on defense order and model choice. Second, our representational analysis covers 11 selected cases and therefore does not establish representational counteraction as a universal mechanism for all conflicts. However, these cases span three model families, both model scales, all three risk dimensions, and different interaction outcomes, providing evidence that the observed pattern extends beyond a single model or defense setting. Finally, CTDR is evaluated on six conflicts, so its generality remains to be established; nevertheless, these results serve as a proof of concept that preserving defense-supporting representational changes can mitigate earlier-objective regression while retaining subsequent-defense gains.

AI USE STATEMENT

This paper was polished for grammar and presentation using ChatGPT. All AI-assisted suggestions were reviewed and verified by the authors, who take full responsibility for the content, experiments, analyses, and conclusions of this work.

REPRODUCIBILITY STATEMENT

We provide the model identifiers, datasets, evaluation metrics, and sequential-composition protocol in the main text. Appendix A describes the defense implementations and primary training configurations, while Appendices B and C provide checkpoint-level risk scores and general-utility results. For the representational analysis and CTDR, we specify the data-split roles, intervention and selection procedures, optimization objectives, and CTDR regularization strength used in the reported experiments. We will release the complete code upon acceptance.

ETHICAL CONSIDERATIONS

This paper studies the interactions between existing defense mechanisms for Large Language Models under sequential deployment, mirroring how LLM providers incrementally patch models in practice. Our work does not introduce new attack methods, does not compromise any deployed system, and does not involve human subjects or real user data. All experiments are conducted on publicly released open-source models using publicly available datasets designed for safety, privacy, and fairness evaluation. The sensitive prompts used to evaluate safety risk are sourced from established benchmarks; we do not create or amplify new harmful content.

The primary goal of this work is defensive: by systematically exposing failure modes in sequential defense composition, we aim to help practitioners and model providers identify and remediate latent security regressions before they can be exploited. We acknowledge that revealing specific defense orderings that produce security regressions (e.g., Catastrophic Collapses) could in principle be misused by adversaries who monitor model update lifecycles. However, consistent with the norm of responsible vulnerability disclosure in security research, we judge that the benefit of informing the research community and model providers of these risks substantially outweighs this concern. To minimize potential misuse, we deliberately frame our findings at the level of interaction patterns and mechanistic explanations rather than providing step-by-step exploitation guides. No institutional IRB/ERB approval was required for this work, as it involves no human participants, no collection of personal data, and no deployment of systems affecting real users.

REFERENCES

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14138–14149, 2024.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025.

Google. gemma-2-2b-it model card. https://huggingface.co/google/gemma-2-2b-it, 2024a.

Google. gemma-7b-it model card. https://huggingface.co/google/gemma-7b-it, 2024b.

Google DeepMind. Introducing the frontier safety framework. https://deepmind.google/blog/introducing-the-frontier-safety-framework/, May 2024. Accessed: 2026-03-31.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, et al. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, 2025.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3923–3954, 2024a.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024b.

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*, 2024c.

Dianqing Liu, Yi Liu, Guoqing Jin, and Zhendong Mao. Mitigating biases in language models via bias unlearning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4160–4178, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Meta Llama Team. Llama-3.2-1b-instruct model card. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, 2024a. Accessed: 2024-09-25.

Meta Llama Team. Meta-llama-3-8b-instruct model card. https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, 2024b.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences. *arXiv preprint arXiv:2510.13900*, 2025.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5356–5371, 2021.

OpenAI. Updated preparedness framework. https://openai.com/index/updating-our-preparedness-framework/, Apr 2025. Accessed: 2026-03-31.

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.

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.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pp. 54911–54941, 2025.

Qwen Team, Alibaba Group. Qwen2.5-1.5b-instruct model card. https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct, 2024a.

Qwen Team, Alibaba Group. Qwen2.5-7b-instruct model card. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct, 2024b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Changsheng Wang, Yihua Zhang, Jinghan Jia, Parikshit Ram, Dennis Wei, Yuguang Yao, Soumyadeep Pal, Nathalie Baracaldo, and Sijia Liu. Invariance makes llm unlearning resilient even to unanticipated downstream fine-tuning. *arXiv preprint arXiv:2506.01339*, 2025.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in llms with continuous attacks. *Advances in Neural Information Processing Systems*, 37:1502–1530, 2024.

Xin Xu, Xunzhi He, Churan Zhi, Ruizhe Chen, Julian McAuley, and Zexue He. Biasfreebench: a benchmark for mitigating bias in large language model responses. *arXiv preprint arXiv:2510.00232*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

A Defense Implementations

We instantiate six defense methods, with two targeting each of safety, privacy, and fairness. These methods cover different update mechanisms, including preference optimization, adversarial training, representation-based unlearning, and parameter editing. Below, we describe their objectives and the datasets used in our evaluation.

Safety defenses. We evaluate preference-based alignment and continuous adversarial training.

- **DPO (Direct Preference Optimization).** DPO (Rafailov et al., 2023) optimizes the relative likelihood of preferred and dispreferred responses using a reference model. For safety alignment, preferred responses follow safety requirements, while dispreferred responses contain unsafe content. We instantiate DPO using preference pairs from PKU-SafeRLHF (Ji et al., 2025).
- **CAT (Continuous Adversarial Training).** CAT (Xhonneux et al., 2024) trains models against adversarial perturbations in the input embedding space. It constructs continuous perturbations that challenge the model's safety behavior and updates the model to maintain safe responses under these perturbations. We instantiate CAT using harmful prompts from HarmBench (Mazeika et al., 2024).

Privacy defenses. We evaluate two unlearning methods on the TOFU forget set (Maini et al., 2024), targeting the influence of information introduced during initialization.

Table 5: Primary training configurations for the six defense methods. Batch size follows the convention of the corresponding implementation.

Defense	Learning Rate	Training Budget	Batch Size
DPO	1×10^{-6}	2 epochs	64
CAT	2×10^{-4}	5 epochs	64
NPO	1×10^{-5}	10 epochs	32
RMU	1×10^{-6}	10 epochs	32
Unbias	2×10^{-5}	early stopping	32
TV	8×10^{-6}	20 epochs	32

Table 6: Baseline risk scores at M_0. Larger values indicate higher risk.

Family	Size	Safety (S)	Privacy (P)	Fairness (F)
Llama	S	0.6535	0.7043	0.3116
	L	0.7046	0.9893	0.3419
Gemma	S	0.5685	0.9841	0.1649
	L	0.6592	0.9963	0.1834
Qwen	S	0.5718	0.1866	0.2602
	L	0.6807	0.8367	0.2935

- **RMU (Representation Misdirection for Unlearning).** RMU (Li et al., 2024b) changes hidden representations associated with content to be forgotten. Its objective steers activations on forget examples toward a fixed random target while encouraging activations on retain examples to remain close to those of a frozen reference model. We apply this representation-based objective using TOFU forget and retain data.
- **NPO (Negative Preference Optimization).** NPO (Zhang et al., 2024) treats forget examples as dispreferred outputs and reduces their likelihood relative to a fixed reference model. This preference-based objective discourages the reproduction of targeted content. We apply NPO to the designated TOFU forget set.

Fairness defenses. We evaluate unlearning-based debiasing and task-vector editing using StereoSet-based data (Nadeem et al., 2021).

- **Unbias.** Unbias (Liu et al., 2025) applies targeted unlearning to stereotypical associations while retaining anti-stereotypical information. Its training objective combines these complementary components to reduce stereotypical preferences. We instantiate the method using StereoSet examples containing stereotypical and anti-stereotypical associations.
- **TV (Task-Vector Editing).** Task-vector editing (Ilharco et al., 2022; Xu et al., 2025) modifies model parameters using a task-specific parameter difference. In our debiasing setting, a bias-related task vector is constructed from the difference between a bias-amplified model and its reference checkpoint. A scaled version of this vector is subtracted from the model being edited to counteract the associated biased behavior. The bias-related training data are derived from StereoSet.

Implementation details. Table 5 summarizes the main training configurations used for the six defenses. Together with the objectives and datasets described above, these settings specify the core implementation of each defense in our evaluation.

B Risk Scores

Table 6 reports the baseline risk scores at M_0, and Tables 7–9 report the checkpoint-level scores for all ordered defense compositions. Here, S, P, and F denote safety, privacy, and fairness, respectively. All scores are oriented so that larger values indicate higher risk.

Table 7: Checkpoint-level risk scores for the Llama family. For each model, *Earlier* denotes the target of D_1 and *Subsequent* denotes the target of D_2.

Primary Defense (D_1)	Subsequent Defense (D_2)	Llama-S				Llama-L			
		Earlier		Subsequent		Earlier		Subsequent	
		M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2
DPO (S)	NPO (P)	0.4670	0.5095	0.7012	0.0562	0.5289	0.4092	0.9881	0.0329
	RMU (P)	0.4670	0.5248	0.7012	0.0657	0.5289	0.4736	0.9881	0.2484
	Unbias (F)	0.4670	0.4175	0.3037	0.0054	0.5289	0.3589	0.3421	0.0212
	TV (F)	0.4670	0.3787	0.3037	0.2586	0.5289	0.1988	0.3421	0.1003
CAT (S)	NPO (P)	0.2954	0.2591	0.1355	0.0993	0.2731	0.1922	0.2180	0.1252
	RMU (P)	0.2954	0.2335	0.1355	0.0731	0.2731	0.2682	0.2180	0.0542
	Unbias (F)	0.2954	0.2525	0.2890	0.0360	0.2731	0.2277	0.3642	0.0975
	TV (F)	0.2954	0.2616	0.2890	0.2036	0.2731	0.2871	0.3642	0.0540
NPO (P)	DPO (S)	0.0748	0.0729	0.5421	0.2591	0.4860	0.2529	0.6898	0.6724
	CAT (S)	0.0748	0.1030	0.5421	0.2979	0.4860	0.1556	0.6898	0.2632
	Unbias (F)	0.0748	0.0734	0.2618	0.0500	0.4860	0.2023	0.3600	0.0189
	TV (F)	0.0748	0.1533	0.2618	0.1595	0.4860	0.1644	0.3600	0.0024
RMU (P)	DPO (S)	0.0722	0.0358	0.6386	0.5553	0.2418	0.0466	0.6840	0.6139
	CAT (S)	0.0722	0.1058	0.6386	0.3894	0.2418	0.0546	0.6840	0.3391
	Unbias (F)	0.0722	0.0349	0.2966	0.0897	0.2418	0.0350	0.3402	0.0190
	TV (F)	0.0722	0.0361	0.2966	0.2034	0.2418	0.0512	0.3402	0.0112
Unbias (F)	DPO (S)	0.1128	0.0563	0.5528	0.4208	0.0927	0.0476	0.6790	0.5272
	CAT (S)	0.1128	0.2321	0.5528	0.3234	0.0927	0.2730	0.6790	0.3375
	NPO (P)	0.1128	0.0331	0.3015	0.0382	0.0927	0.0373	0.9487	0.0346
	RMU (P)	0.1128	0.1118	0.3015	0.0468	0.0927	0.0548	0.9487	0.0325
TV (F)	DPO (S)	0.1699	0.1828	0.5710	0.4637	0.0644	0.0641	0.5941	0.4455
	CAT (S)	0.1699	0.2270	0.5710	0.3746	0.0644	0.1605	0.5941	0.2781
	NPO (P)	0.1699	0.1714	0.1262	0.0356	0.0644	0.0174	0.3262	0.0435
	RMU (P)	0.1699	0.1719	0.1262	0.0375	0.0644	0.0272	0.3262	0.0589

For an ordered composition $D_1 \rightarrow D_2$, the target dimension of each defense is shown in parentheses after its name. Under each model, the *Earlier* columns report $S_{R_1}(M_1)$ and $S_{R_1}(M_2)$, while the *Subsequent* columns report $S_{R_2}(M_1)$ and $S_{R_2}(M_2)$. We retain $S_{R_2}(M_1)$ because the gain of the subsequent defense is defined from the checkpoint at which that defense is applied, $G_2 = S_{R_2}(M_1) - S_{R_2}(M_2)$. Comparing M_2 only with M_0 would instead measure the net change from the initial model and would not isolate the effect of D_2.

Because all compositions for a given model share the same M_0, the baseline scores are reported once in Table 6 rather than repeated in every composition row. All displayed scores are rounded to four decimal places for readability. All defense gains, regressions, RRR values, and interaction classifications are computed from the corresponding unrounded checkpoint-level scores.

C General Utility Evaluation

To examine whether the observed defense interactions could be explained by broad capability degradation, we additionally evaluate MMLU accuracy for all sequential defense compositions. As shown in Table 10, most compositions remain within a few percentage points of their corresponding undefended baselines. This suggests that the risk regressions reported in Section 2.3 are not generally attributable to an overall collapse in model capability. Some compositions, particularly those involving Task Vector (TV) editing, exhibit somewhat larger decreases in MMLU accuracy, indicating that general-utility degradation can occur in particular settings and should be considered separately from sequential defense compatibility.

Table 8: Checkpoint-level risk scores for the Gemma family. For each model, *Earlier* denotes the target of D_1 and *Subsequent* denotes the target of D_2.

Primary Defense (D_1)	Subsequent Defense (D_2)	Gemma-S				Gemma-L			
		Earlier		Subsequent		Earlier		Subsequent	
		M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2
DPO (S)	NPO (P)	0.4051	0.5190	0.9761	0.0512	0.2673	0.1444	0.9919	0.0306
	RMU (P)	0.4051	0.4266	0.9761	0.1064	0.2673	0.2706	0.9919	0.2220
	Unbias (F)	0.4051	0.5355	0.2080	0.0287	0.2673	0.2880	0.1039	0.0406
	TV (F)	0.4051	0.4876	0.2080	0.0579	0.2673	0.1139	0.1039	0.0013
CAT (S)	NPO (P)	0.3127	0.3358	0.1477	0.0876	0.2913	0.3729	0.1427	0.0306
	RMU (P)	0.3127	0.2880	0.1477	0.0624	0.2913	0.3053	0.1427	0.0525
	Unbias (F)	0.3127	0.2203	0.1601	0.0459	0.2913	0.4417	0.1472	0.0480
	TV (F)	0.3127	0.2451	0.1601	0.0259	0.2913	0.1634	0.1472	0.0395
NPO (P)	DPO (S)	0.0306	0.0306	0.8078	0.6081	0.0306	0.0306	0.5215	0.3837
	CAT (S)	0.0306	0.0657	0.8078	0.2401	0.0306	0.0458	0.5215	0.3507
	Unbias (F)	0.0306	0.0306	0.1624	0.0867	0.0306	0.0306	0.1587	0.0314
	TV (F)	0.0306	0.0306	0.1624	0.1616	0.0306	0.0306	0.1587	0.0224
RMU (P)	DPO (S)	0.1695	0.1875	0.6238	0.4026	0.1890	0.0611	0.6823	0.2756
	CAT (S)	0.1695	0.1245	0.6238	0.1848	0.1890	0.0431	0.6823	0.4010
	Unbias (F)	0.1695	0.0310	0.1049	0.0133	0.1890	0.0357	0.1243	0.0599
	TV (F)	0.1695	0.0306	0.1049	0.0797	0.1890	0.0306	0.1243	0.0518
Unbias (F)	DPO (S)	0.0142	0.0750	0.7343	0.4563	0.0502	0.0103	0.6370	0.2797
	CAT (S)	0.0142	0.1102	0.7343	0.2434	0.0502	0.0903	0.6370	0.5173
	NPO (P)	0.0142	0.0404	0.4603	0.0306	0.0502	0.0081	0.9799	0.0306
	RMU (P)	0.0142	0.0374	0.4603	0.0315	0.0502	0.0789	0.9799	0.1528
TV (F)	DPO (S)	0.0841	0.0487	0.6328	0.3705	0.0291	0.0535	0.5347	0.1667
	CAT (S)	0.0841	0.1306	0.6328	0.2781	0.0291	0.0581	0.5347	0.4975
	NPO (P)	0.0841	0.1092	0.3052	0.0306	0.0291	0.0290	0.2305	0.0306
	RMU (P)	0.0841	0.0124	0.3052	0.0306	0.0291	0.0891	0.2305	0.0306

Table 9: Checkpoint-level risk scores for the Qwen family. For each model, *Earlier* denotes the target of D_1 and *Subsequent* denotes the target of D_2.

Primary Defense (D_1)	Subsequent Defense (D_2)	Qwen-S				Qwen-L			
		Earlier		Subsequent		Earlier		Subsequent	
		M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2
DPO (S)	NPO (P)	0.4472	0.3878	0.1836	0.0770	0.5528	0.5446	0.8171	0.8153
	RMU (P)	0.4472	0.4307	0.1836	0.1243	0.5528	0.5264	0.8171	0.8049
	Unbias (F)	0.4472	0.4290	0.2605	0.0007	0.5528	0.4282	0.2936	0.0481
	TV (F)	0.4472	0.2657	0.2605	0.1532	0.5528	0.1865	0.2936	0.1911
CAT (S)	NPO (P)	0.1436	0.1799	0.0834	0.0300	0.0058	0.0033	0.0927	0.0902
	RMU (P)	0.1436	0.1139	0.0834	0.0727	0.0058	0.0041	0.0927	0.0926
	Unbias (F)	0.1436	0.1601	0.2610	0.0341	0.0058	0.0074	0.3240	0.0655
	TV (F)	0.1436	0.2723	0.2610	0.1245	0.0058	0.2871	0.3240	0.0547
NPO (P)	DPO (S)	0.0415	0.0302	0.5635	0.4744	0.0309	0.0312	0.6411	0.4373
	CAT (S)	0.0415	0.0402	0.5635	0.1064	0.0309	0.0673	0.6411	0.0033
	Unbias (F)	0.0415	0.0302	0.2413	0.0354	0.0309	0.0304	0.2952	0.0447
	TV (F)	0.0415	0.0300	0.2413	0.1299	0.0309	0.0300	0.2952	0.1711
RMU (P)	DPO (S)	0.1028	0.1042	0.5792	0.4505	0.1774	0.1314	0.7104	0.5182
	CAT (S)	0.1028	0.0730	0.5792	0.1205	0.1774	0.0916	0.7104	0.0017
	Unbias (F)	0.1028	0.0710	0.2515	0.0250	0.1774	0.1087	0.3058	0.0435
	TV (F)	0.1028	0.0300	0.2515	0.1188	0.1774	0.0300	0.3058	0.1979
Unbias (F)	DPO (S)	0.0003	0.0129	0.6147	0.4109	0.0323	0.0118	0.6939	0.4579
	CAT (S)	0.0003	0.1717	0.6147	0.1304	0.0323	0.1697	0.6939	0.0066
	NPO (P)	0.0003	0.0035	0.1307	0.0467	0.0323	0.0847	0.6076	0.0763
	RMU (P)	0.0003	0.0073	0.1307	0.0668	0.0323	0.0510	0.6076	0.0378
TV (F)	DPO (S)	0.1162	0.1188	0.4893	0.3325	0.1017	0.0996	0.4761	0.3135
	CAT (S)	0.1162	0.2039	0.4893	0.1081	0.1017	0.1436	0.4761	0.0000
	NPO (P)	0.1162	0.1204	0.1300	0.0300	0.1017	0.1129	0.1300	0.0300
	RMU (P)	0.1162	0.1031	0.1300	0.0300	0.1017	0.0068	0.1300	0.0300

Table 10: MMLU accuracy (%) across all sequential defense compositions, reported as mean ± std. Most compositions remain close to the corresponding undefended baseline, suggesting no general capability collapse.

Primary Defense (D_1)	**Subsequent Defense** (D_2)	**Llama**		**Gemma**		**Qwen**	
		-S	**-L**	**-S**	**-L**	**-S**	**-L**
No Defense	–	$46.03_{\pm 0.41}$	$66.10_{\pm 0.38}$	$56.44_{\pm 0.40}$	$51.70_{\pm 0.39}$	$59.75_{\pm 0.39}$	$72.33_{\pm 0.36}$
		Fairness-First					
Unbias (Fairness Stage)	+ RMU (Privacy)	$44.05_{\pm 0.41}$	$65.18_{\pm 0.39}$	$54.45_{\pm 0.40}$	$51.24_{\pm 0.40}$	$57.63_{\pm 0.40}$	$71.73_{\pm 0.38}$
	+ NPO (Privacy)	$45.39_{\pm 0.41}$	$65.35_{\pm 0.38}$	$56.02_{\pm 0.40}$	$51.70_{\pm 0.39}$	$59.69_{\pm 0.39}$	$70.89_{\pm 0.40}$
	+ DPO (Safety)	$45.85_{\pm 0.41}$	$65.63_{\pm 0.38}$	$56.27_{\pm 0.40}$	$50.75_{\pm 0.39}$	$59.59_{\pm 0.39}$	$72.16_{\pm 0.36}$
	+ CAT (Safety)	$43.71_{\pm 0.41}$	$64.91_{\pm 0.39}$	$53.73_{\pm 0.40}$	$50.13_{\pm 0.38}$	$59.28_{\pm 0.39}$	$71.19_{\pm 0.36}$
Task Vector (TV) (Fairness Stage)	+ RMU (Privacy)	$43.95_{\pm 0.40}$	$63.86_{\pm 0.40}$	$54.78_{\pm 0.41}$	$50.31_{\pm 0.40}$	$58.44_{\pm 0.41}$	$70.40_{\pm 0.37}$
	+ NPO (Privacy)	$44.20_{\pm 0.41}$	$64.66_{\pm 0.41}$	$54.94_{\pm 0.41}$	$50.52_{\pm 0.39}$	$57.55_{\pm 0.40}$	$71.58_{\pm 0.41}$
	+ DPO (Safety)	$44.12_{\pm 0.41}$	$64.64_{\pm 0.41}$	$54.46_{\pm 0.41}$	$51.04_{\pm 0.39}$	$56.92_{\pm 0.41}$	$71.40_{\pm 0.41}$
	+ CAT (Safety)	$43.48_{\pm 0.41}$	$63.81_{\pm 0.39}$	$53.78_{\pm 0.40}$	$49.74_{\pm 0.40}$	$58.63_{\pm 0.39}$	$70.38_{\pm 0.37}$
		Privacy-First					
NPO (Privacy Stage)	+ Unbias (Fairness)	$45.82_{\pm 0.41}$	$65.03_{\pm 0.38}$	$56.65_{\pm 0.40}$	$51.31_{\pm 0.40}$	$58.42_{\pm 0.40}$	$72.58_{\pm 0.36}$
	+ TV (Fairness)	$44.13_{\pm 0.41}$	$63.86_{\pm 0.38}$	$54.43_{\pm 0.41}$	$50.01_{\pm 0.39}$	$58.35_{\pm 0.40}$	$71.09_{\pm 0.39}$
	+ DPO (Safety)	$44.50_{\pm 0.41}$	$65.40_{\pm 0.38}$	$56.50_{\pm 0.40}$	$51.84_{\pm 0.41}$	$59.76_{\pm 0.39}$	$72.55_{\pm 0.36}$
	+ CAT (Safety)	$43.78_{\pm 0.41}$	$64.27_{\pm 0.39}$	$53.75_{\pm 0.40}$	$51.02_{\pm 0.40}$	$59.31_{\pm 0.39}$	$71.62_{\pm 0.36}$
RMU (Privacy Stage)	+ Unbias (Fairness)	$45.21_{\pm 0.41}$	$64.25_{\pm 0.38}$	$54.99_{\pm 0.40}$	$51.11_{\pm 0.38}$	$58.56_{\pm 0.40}$	$69.62_{\pm 0.37}$
	+ TV (Fairness)	$44.39_{\pm 0.40}$	$63.57_{\pm 0.39}$	$53.92_{\pm 0.41}$	$50.88_{\pm 0.40}$	$57.49_{\pm 0.41}$	$70.75_{\pm 0.41}$
	+ DPO (Safety)	$44.00_{\pm 0.41}$	$64.68_{\pm 0.38}$	$55.96_{\pm 0.40}$	$50.92_{\pm 0.41}$	$59.00_{\pm 0.40}$	$70.60_{\pm 0.37}$
	+ CAT (Safety)	$43.23_{\pm 0.41}$	$64.41_{\pm 0.39}$	$53.80_{\pm 0.40}$	$50.08_{\pm 0.38}$	$58.70_{\pm 0.39}$	$70.87_{\pm 0.36}$
		Safety-First					
DPO (Safety Stage)	+ Unbias (Fairness)	$45.41_{\pm 0.38}$	$64.58_{\pm 0.38}$	$53.42_{\pm 0.41}$	$51.22_{\pm 0.40}$	$59.46_{\pm 0.39}$	$72.37_{\pm 0.36}$
	+ TV (Fairness)	$44.07_{\pm 0.41}$	$63.01_{\pm 0.40}$	$53.95_{\pm 0.41}$	$50.26_{\pm 0.41}$	$57.34_{\pm 0.41}$	$71.01_{\pm 0.40}$
	+ NPO (Privacy)	$44.64_{\pm 0.41}$	$64.73_{\pm 0.38}$	$56.05_{\pm 0.40}$	$51.38_{\pm 0.40}$	$59.77_{\pm 0.39}$	$72.15_{\pm 0.37}$
	+ RMU (Privacy)	$43.90_{\pm 0.41}$	$63.56_{\pm 0.39}$	$55.08_{\pm 0.40}$	$51.34_{\pm 0.39}$	$57.71_{\pm 0.40}$	$71.94_{\pm 0.37}$
CAT (Safety Stage)	+ Unbias (Fairness)	$44.53_{\pm 0.41}$	$65.79_{\pm 0.39}$	$54.26_{\pm 0.40}$	$51.00_{\pm 0.40}$	$59.30_{\pm 0.39}$	$70.99_{\pm 0.36}$
	+ TV (Fairness)	$43.81_{\pm 0.37}$	$64.65_{\pm 0.36}$	$53.65_{\pm 0.36}$	$49.59_{\pm 0.41}$	$57.45_{\pm 0.41}$	$69.89_{\pm 0.37}$
	+ NPO (Privacy)	$44.80_{\pm 0.41}$	$64.76_{\pm 0.39}$	$54.33_{\pm 0.40}$	$51.61_{\pm 0.40}$	$59.14_{\pm 0.39}$	$71.90_{\pm 0.37}$
	+ RMU (Privacy)	$43.89_{\pm 0.41}$	$63.72_{\pm 0.40}$	$54.82_{\pm 0.40}$	$51.44_{\pm 0.41}$	$58.21_{\pm 0.40}$	$71.80_{\pm 0.37}$