LARGER IS NOT ALWAYS BETTER: PERTURBATION BUDGETS IN ADVERSARIAL TRAINING FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-space adversarial training offers a scalable way to harden large language models against jailbreaks by optimizing adversarial perturbations in continuous representation space. A natural design intuition, inherited from computer vision, is that giving the training adversary a larger perturbation budget should expose the model to stronger attacks and thereby improve robustness. In learned LLM representation spaces, however, perturbation scales lack a natural interpretation, and their effects may differ across robustness and utility evaluations. We test how far this intuition holds through controlled perturbation-budget sweeps across configurations, evaluating the resulting models on jailbreak attacks, general capabilities, and benign compliance. We find no universal monotonic robustness trend: increasing the budget can improve robustness against some attacks, leave others unchanged, or even make robustness worse, depending on the attack and training configuration. The utility cost is equally uneven. Benign compliance can collapse while capability benchmarks remain comparatively stable, whereas in other settings capability degrades substantially with much smaller changes in benign compliance. Moreover, sufficiently large budgets can produce strictly worse robustness–utility trade-offs than smaller ones. To understand why budget alone is an incomplete description of training, we measure the perturbations actually realized by the inner optimizer. Larger configured budgets do produce larger perturbations, but with decreasing budget utilization, while the perturbations remain local to their original token embeddings. We further ask whether the dominant behavioral failure mode, over-refusal, can be repaired. Targeted supervision substantially improves behavior on the supervised over-refusal distribution, yet does not reliably recover broader benign compliance and consistently harms adversarial robustness. Overall, our results highlight both the promise and the challenges of training against stronger latent-space adversaries: larger perturbation budgets can improve robustness, but future methods must use that additional budget effectively, generalize robustness across attack families, remain stable across training runs, and avoid behavioral costs such as over-refusal.

1 INTRODUCTION

Adversarial training improves the robustness of large language models (LLMs) to jailbreak attacks by exposing them to adversarially perturbed inputs during training (Mazeika et al., 2024). Latent-space methods make this procedure more tractable by optimizing perturbations over continuous learned representations rather than directly searching the discrete space of text inputs (Xhonneux et al., 2024; Sheshadri et al., 2024; Dékány et al., 2026). A key design choice in such methods is the perturbation budget assigned to the training adversary, typically enforced by constraining the perturbation to lie within an ℓ_p ball of radius ϵ around the original input.

There are several reasons to expect larger-radius training to improve robustness. Increasing ϵ enlarges the adversary's feasible set, thereby exposing the model to stronger adversarial examples during training. This intuition is well established in adversarial training for computer vision (Madry et al., 2017; Li & Spratling, 2023). Whether the same relationship holds reliably for LLMs, however, is less clear. Robustness gains may not transfer uniformly across jailbreak attacks that target different inputs or operate under different attack budgets. Utility is also multidimensional for aligned LLMs, so the cost of a larger radius may differ across task capability and benign behavior. More-

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.

over, the interpretation of the radius itself is less straightforward in learned representation spaces, where distances lack the external perceptual meaning and natural scale available for image-space perturbations (Madry et al., 2017).

Prior work provides partial evidence. Continuous Adversarial Training (CAT) perturbs input embeddings, while Latent Adversarial Training (LAT) additionally perturbs intermediate model activations (Xhonneux et al., 2024; Sheshadri et al., 2024). The original CAT work reports a robustness–utility relationship across a radius ablation for its preference-optimization variant, CAPO, but varies the radius jointly with another hyperparameter and uses a single base model (Xhonneux et al., 2024). Later studies investigate perturbation budgets and related inner-loop parameters in other latent-space adversarial-training settings (Dékány et al., 2026; Fu & Wang, 2026), but do not isolate CAT's radius across multiple configurations, while jointly measuring realized perturbations and evaluating resulting checkpoints across attack families and utility measures.

We therefore study three questions. First, and most importantly, how does increasing the perturbation radius affect robustness, and what utility costs does it incur in latent-space adversarial training? Second, what optimization or behavioral challenges emerge as the radius increases? Third, can the most prominent behavioral cost—over-refusal on benign requests—be mitigated through targeted supervision? We address these questions through controlled perturbation-radius sweeps using CAT across three model–recipe configurations, complemented by LAT sweeps. We evaluate each resulting checkpoint against up to four jailbreak attacks, measure task capability and benign compliance, analyze the perturbations realized by the inner-loop adversary during training, and perform targeted over-refusal training to test whether benign compliance can be recovered by training on over-refusal data. We obtain three main findings:

(1) **Robustness varies non-monotonically with the perturbation budget.** Increasing the budget initially improves robustness, but these gains diminish and can reverse at larger budgets. The location and magnitude of this peak depend on both the attack and the model–recipe configuration, so no single budget induces a consistent robustness ordering across settings. Meanwhile, benign compliance can deteriorate substantially before general task capability, and sufficiently large budgets can become dominated, incurring additional utility loss without compensating robustness gains.

(2) **Increasing the perturbation budget changes both the realized adversary and training variability.** Realized perturbation norms grow monotonically with the budget, but less than proportionally, while remaining local to their corresponding token embeddings. Several intermediate and large-budget settings also exhibit increased between-seed variability, particularly in benign compliance, although this pattern is not systematic across configurations.

(3) **Targeted over-refusal supervision generalizes within the training distribution but not reliably across benign-compliance datasets.** The intervention improves benign compliance on held-out examples from the same dataset used for supervision, but these gains do not consistently transfer to independent benign-compliance datasets and are accompanied by reduced adversarial robustness.

2 Background and Related Work

2.1 Latent-Space Adversarial Training

Adversarial training improves model robustness by training against inputs constructed by an adversarial inner optimization (Madry et al., 2017). For LLMs, this optimization can search directly over discrete token sequences, as in R2D2 (Mazeika et al., 2024), but such a search is computationally expensive. Latent-space methods instead optimize perturbations in differentiable representation spaces. CAT perturbs input-token embeddings (Xhonneux et al., 2024), while LAT perturbs both input embeddings and intermediate activations (Sheshadri et al., 2024); related methods also perturb activations or combine continuous and discrete adversarial examples (Dékány et al., 2026; Fu & Wang, 2026). We study CAT in greater detail because its input-embedding perturbations have a directly measurable geometry, and complement it with LAT to test whether the robustness and utility patterns extend to another latent-space adversarial-training setup.

Let $E \in \mathbb{R}^{|V| \times d}$ denote the token-embedding matrix. CAT independently adds a perturbation $\delta_i \in \mathbb{R}^d$ to each perturbed input position i, subject to

$$\|\delta_i\|_2 \leq \epsilon. \tag{1}$$

Perturbations are applied to non-response positions, including both user-content and template tokens. Given a harmful prompt and a target continuation, CAT's inner optimization searches within this constraint for a perturbation that increases the likelihood of the harmful continuation. The outer optimization then trains the model to produce a safe continuation and suppress the harmful one under the perturbed input, while also optimizing utility on unperturbed benign data (Xhonneux et al., 2024). In our experiments, we vary the radius of the inner perturbation while keeping the remaining training objective fixed within each configuration.

2.2 Why a Larger Perturbation Radius Might Help

The radius ϵ defines the feasible perturbation set available to CAT's inner optimizer. In the released implementation, it is derived from a dimensionless multiplier s and the mean norm of the base model's token embeddings:

$$\epsilon = s \cdot \|E\|_{\text{mean}}, \qquad \|E\|_{\text{mean}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \|E_v\|_2. \tag{2}$$

This scaling accounts for differences in embedding scale between models, but it does not give ϵ the same interpretation as a perturbation budget in domains such as images. Embedding coordinates have no external unit or perceptual threshold, so two models with the same relative radius do not necessarily respond to that radius in the same way.

Nevertheless, larger-radius training is well motivated: a larger feasible set contains every perturbation available at a smaller one. Xhonneux et al. (2024) report a robustness–utility analysis, where larger radii are associated with greater GCG robustness and lower MMLU performance, while Fu & Wang (2026) derive a radius-dependent robustness bound for linear transformers. However, the CAPO ablation jointly varies radius and IPO coefficient on one base model and does not report realized perturbations. MixAT (Dékány et al., 2026) likewise varies radius together with inner steps and step size. Fu & Wang (2026)'s bound also depends on embedding-matrix properties, motivating model-dependent effects.

Two properties of the training procedure complicate the expectation. We measure both. First, ϵ defines the feasible perturbation rather than the distance ultimately reached by CAT's finite inner optimization. We therefore distinguish the configured radius from the realized perturbation distance:

$$D = \mathbb{E}[\|\delta^*\|_2], \qquad \rho = \frac{D}{\epsilon}, \tag{3}$$

where δ^* denotes the lowest-loss inner-loop iterate selected for training, and the expectation is taken over perturbed positions and training steps. The utilization ratio ρ expresses the realized distance as a fraction of the available radius. In this paper, ϵ is the experimental variable, while D and ρ diagnose how changes in ϵ affect the perturbations actually used during training. Second, the perturbed embeddings sit in a space that also contains the vocabulary embeddings, so a larger feasible region raises the question of whether the intervention remains a local perturbation of the original token.

Assessing the cost of increasing the radius also requires distinguishing between different aspects of model utility. Likelihood-based capability benchmarks can remain high even when an aligned model increasingly refuses benign requests during generation. CAT itself has been shown to exhibit over-refusal on harmless prompts (Xhonneux et al., 2024), while subsequent safety-evaluation work has emphasized evaluating over-refusal separately from general capability (Beyer et al., 2026). We therefore evaluate both to ask how task capability and benign compliance respond as the perturbation radius increases across the evaluated latent-space adversarial-training setups.

3 Experimental Design

CAT perturbation-radius sweeps. Within each model–recipe configuration, we vary CAT's configured radius ϵ while holding training data, adaptation, epochs, outer optimizer, $K = 10$ inner iterations, and relative update scale fixed. Dimension-normalized inner updates have ℓ_2 norm 0.1ϵ. Thus, increasing ϵ expands the feasible region without changing update size relative to that region. Each radius uses three independent training seeds. Comparisons are made within each radius sweep. Cross-model comparisons are descriptive because embedding geometries differ. Full training implementation details are provided in Appendix A.1.

Model–recipe configurations and radius coverage. We consider three model–recipe configurations across two base-model families: **Zephyr** using `HuggingFaceH4/zephyr-7b-beta` (Tunstall et al., 2023) with its released CAT training recipe; **Llama 3 (Llama 2 recipe)** using `Meta-Llama-3-8B-Instruct` (Grattafiori et al., 2024) with CAT's Llama 2 recipe; and **Llama 3 (MixAT recipe)** using the same base model with MixAT's Llama 3 recipe (Dékány et al., 2026). The original CAT work does not document a Llama 3 training configuration, so the two Llama 3 settings test sensitivity to recipe transfer rather than represent independent model families. All configurations use HarmBench adversarial training data (Mazeika et al., 2024), UltraChat utility data (Ding et al., 2023), and LoRA adaptation (Hu et al., 2021). We vary the configuration multiplier s geometrically, doubling s at each step. Zephyr and Llama 3 (MixAT recipe) are evaluated at four radii each, while the Llama 2-recipe configuration interleaves two four-point sweeps into eight settings over a denser range. Overall, we train 48 checkpoints across 16 positive-radius settings. Unmodified models ($\epsilon = 0$, no adversarial training) serve only as reference conditions and are excluded from all radius-dependent and dominance analyses. Full hyperparameters are reported in Table 1.

LAT sweeps. We additionally evaluate LAT on Llama 3, using its released configuration, which perturbs the input embeddings and four intermediate layers using $K = 16$ AdamW inner iterations at $\epsilon = 6$ and inner learning rate 10^{-3}. We sweep $\epsilon \in \{6, 12, 24, 48\}$ under two learning-rate conventions: *original-LR* scaling, which scales the released learning-rate proportionally with ϵ, and *10%-step* scaling, which sets it so that the first AdamW update has ℓ_2 norm $\approx 0.1\epsilon$, matching CAT's relative step. Each setting uses three seeds. Inner learning rates and full training details are in Appendix A.1.

Adversarial robustness. We evaluate robustness using four jailbreak attacks and report attack success rate (ASR, lower is better). GCG (Zou et al., 2023) is an adaptive discrete suffix attack. AutoDAN (Liu et al., 2024) adaptively searches over natural-language jailbreak prompts. Soft-Prompts (Schwinn et al., 2024) is an adaptive continuous attack that optimizes additional input embeddings, so it acts at the same model-input level as CAT, although not geometrically equivalent. Inpainting (Lüdke et al., 2025), by contrast, is a high-budget transfer attack evaluated best-of-$1{,}024$ rather than through per-checkpoint adaptive optimization. The attacks differ in inputs, budgets, procedures, and judges, so ASRs are compared only within attack families. Full protocols are in Appendix A.2.

Utility and statistical reporting. We evaluate task capability using MMLU (Hendrycks et al., 2020), ARC-Easy, and ARC-Challenge (Clark et al., 2018), with multiple-choice likelihood scoring. We evaluate benign behavioral utility using Harmless (Xhonneux et al., 2024) and XSTest (Röttger et al., 2024). XSTest reports both safe compliance and unsafe refusal. Following the MixAT protocol, capability benchmarks are evaluated without a chat template, whereas behavioral evaluations use it. Higher accuracy is better. For each radius, we report the mean and sample standard deviation across three independently trained checkpoints. Attack seeds are fixed across training seeds, so reported variation reflects checkpoint variation.

4 Does Training with a Larger Radius Help?

4.1 Robustness Across Attack Families

Increasing the perturbation radius does not induce a consistent robustness ordering across attacks, and LAT shows that even the direction of this response can depend on the training setup. Relative to the unmodified $\epsilon = 0$ reference models, the positive-radius CAT checkpoints generally show substantial robustness gains, particularly against GCG and AutoDAN, and all LAT checkpoints also have lower mean GCG and AutoDAN ASR than the unmodified Llama 3 model. The question addressed here is therefore not whether adversarial training improves robustness, but whether increasing the radius improves robustness further. Figure 1(a–b) shows the two attacks that distinguish CAT radius settings across most of the evaluated range.

GCG responds differently across configurations. Both Llama 3 configurations reach their lowest mean ASR at intermediate radii, whereas Zephyr reaches its lowest value at the largest radius. Several neighboring settings have seed variation comparable to their differences in mean ASR, so we do not interpret the sweep as identifying a precise optimal radius. Thus, GCG does not induce a common monotonic ordering across configurations.

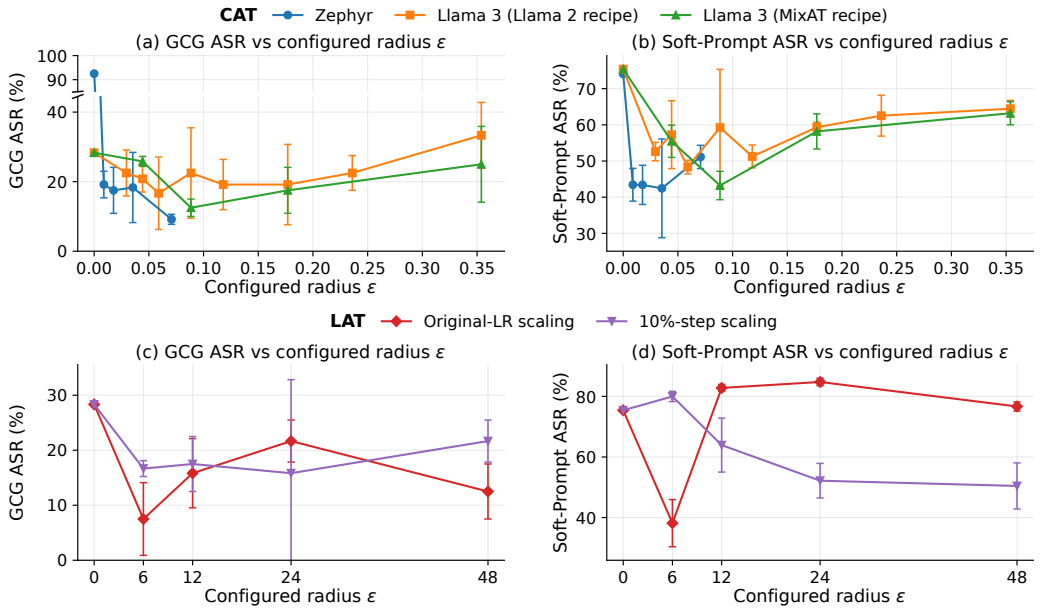

Figure 1: **Larger perturbation radii do not provide a consistent ordering of adversarial robustness.** (a–b) CAT: GCG ASR is non-monotonic for Llama 3, while Zephyr is most robust at the largest radius; for Soft-Prompts, the largest radius is never most robust and has always higher mean ASR than the smallest. (c–d) LAT: GCG shows no monotonic trend; Soft-Prompt robustness improves along the 10%-step sweep, while original-LR ASR mostly exceeds the baseline. Points show three-seed means ± one standard deviation. $\epsilon = 0$ is the unmodified model and so has no error bars. Lower ASR is better. Lines are visual guides. ASRs are compared only within attack families.

The Soft-Prompt attack gives the clearest counterexample to the idea that larger radii should consistently produce greater robustness. Across all three configurations, the largest radius has higher mean ASR than the smallest and is never the most robust setting. Its non-monotonic response is informative because Soft-Prompts acts at the same input level as CAT through a different mechanism: it optimizes an additional, unconstrained block of continuous input vectors. Section 5.2 compares the displacements realized during CAT training and Soft-Prompt evaluation and rules out a simple magnitude explanation for this pattern.

The remaining attacks are less discriminative (Appendix B.3). AutoDAN shows robustness gains but quickly reaches a floor, while Inpainting remains near its ceiling where evaluated. Because Inpainting is a transfer-style evaluation, we interpret this as little sensitivity to ϵ under that evaluation rather than equivalent robustness more generally. Four Llama 3 (Llama 2 recipe) Inpainting evaluations are unavailable.

LAT provides a complementary test of whether CAT's robustness patterns extend to another latent-space adversarial-training method. GCG robustness remains better than the unmodified Llama 3 model throughout both sweeps but does not improve systematically. Soft-Prompts, instead, distinguishes the two sweeps. Under the original-LR scaling, robustness is best at the smallest setting and then becomes worse than the unmodified model. Under 10%-step, robustness monotonically improves across the sweep. Thus, LAT supports the CAT finding that robustness gains are attack-specific, but shows that even the direction of the response to scaling depends on the training setup.

4.2 Task Capability and Benign Compliance

The utility cost of a larger perturbation depends strongly on which component of utility is measured. Complete results are in Appendix B.2. Figure 2(a–b) compares MMLU and Harmless across the three CAT sweeps. The clearest separation appears for Llama 3 with the Llama 2 recipe. Relative to its smallest-radius checkpoint ($s = 0.05$, $\epsilon = 0.02949$), increasing the radius to $s = 0.15$ ($\epsilon = 0.08848$) reduces the mean Harmless successful-response rate from 95.83 to 45.83, a 52% rel-

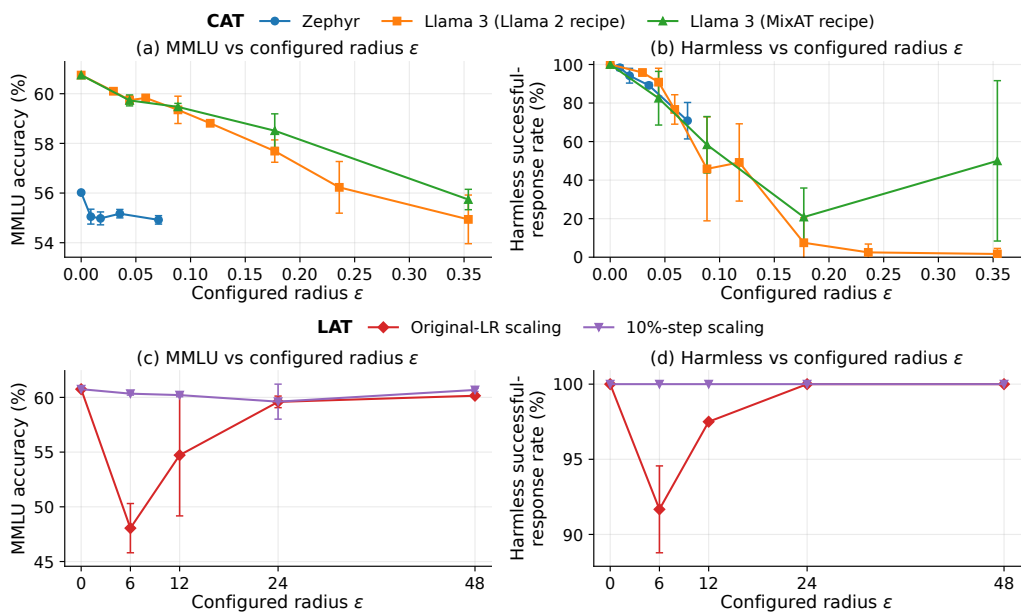

Figure 2: **The utility component most affected differs across methods.** (a–b) CAT: MMLU and Harmless against configured radius ϵ. Harmless declines steeply while MMLU declines gradually. (c–d) LAT: MMLU and Harmless for LAT. Harmless remains at or near ceiling. $\epsilon = 0$ denotes the unmodified base-model reference and has no error bar. Positive-radius points show three-seed means and error bars show one standard deviation.

ative decline. Over the same change, MMLU falls from 60.10 to 59.35 (1%), while ARC-Easy falls from 80.95 to 79.45 (2%). At $s = 0.3$ ($\epsilon = 0.17695$), Harmless falls further to 7.50, corresponding to a 92% decline relative to the smallest-radius setting, whereas MMLU and ARC-Easy decline by only 4% and 7%, respectively. ARC-Challenge shows a similarly gradual decline. Zephyr shows the same qualitative separation, while the largest MixAT radius is highly variable. XSTest safe compliance provides additional evidence that benign response behavior changes with ϵ, but more noisily and non-monotonically, with substantial between-seed variation. Its prefix-based refusal classifier disagrees with an auxiliary judge in a checkpoint-dependent way, and with Harmless at two settings where the classifier also disagrees with the auxiliary judge, so we do not assume that different benign-compliance measures induce the same ordering. Overall, these results show that capability benchmarks alone can substantially understate the behavioral utility cost of increasing CAT's radius.

LAT shows a different utility response. Under the original-LR scaling, the smallest setting substantially degrades task capability relative to the unmodified model, while Harmless changes much less. As ϵ increases, task capability progressively recovers to baseline performance. So, within this sweep, moving toward the larger settings is associated with lower utility cost, rather than the increasing behavior observed across the larger-radius CAT settings. In the 10%-step sweep, both utility components remain stable, so a larger configured budget does not need to incur a larger measured utility cost. The common finding is therefore not that larger-budget adversarial training degrades the same utility component, but that utility responses are method- and configuration-dependent. The utility cost of larger perturbations is not uniform: CAT primarily exposes a benign-compliance cost, whereas the reproduced LAT original-LR sweep primarily affects task capability.

4.3 Joint Robustness–Utility Comparisons

Beyond part of the evaluated CAT range, additional benign-compliance loss does not buy greater measured robustness. We therefore analyze CAT's utility and robustness jointly using dominance among the evaluated radius settings (Figure 3). For a given utility metric and attack, setting A dominates setting B if A has at least as much utility and no greater ASR, with at least one strict improvement. We compute this relation separately for each model–recipe configuration and

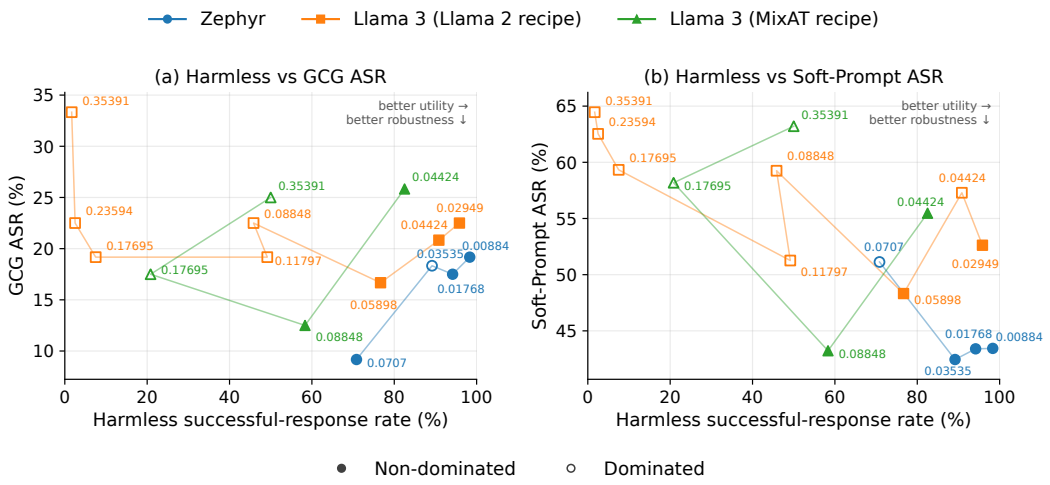

Figure 3: **Increasing the CAT radius can induce substantial behavioral utility costs without continuing to improve measured robustness.** Harmless against (a) GCG and (b) Soft-Prompt ASR. Higher Harmless and lower ASR are better. Points are labeled by configured radius. Filled markers denote observed non-dominated settings within each model–recipe configuration and robustness–utility pairing. Lines connect successive radii only to indicate the sweep trajectory.

each robustness–utility pairing. Since only a finite set of radii is evaluated, this analysis identifies observed non-dominated settings rather than a continuous Pareto frontier, and specifies no preference weighting among them: distinct non-dominated settings represent distinct robustness–utility trade-offs, not a ranking. We focus on Harmless and MMLU for utility and GCG and Soft-Prompts for robustness, as these attacks distinguish checkpoints across most of the swept range.

At small and intermediate radii, several settings remain non-dominated, corresponding to distinct observed robustness–utility trade-offs. For Llama 3, sufficiently large settings become dominated as utility deteriorates without compensating robustness gains. Zephyr remains more attack-dependent. XSTest produces a few exceptions, but substantial seed variation and classifier limitations make these fine-grained differences difficult to interpret.

5 What Challenges Arise at Larger Radii?

The preceding section shows that larger radii do not consistently improve robustness and can carry substantial behavioral costs. We next examine three properties of larger-radius training that qualify these results: the perturbation the inner optimizer actually produces, its position relative to the discrete vocabulary, and between-seed variability.

5.1 Configured Radius and Realized Perturbation

Larger CAT radii produce larger realized perturbations, but the inner optimizer uses a decreasing fraction of the available budget, even with a fixed relative update scale. A natural concern is that the largest-ϵ checkpoints might realize smaller perturbations despite their larger configured budgets. Figure 4(a) shows the opposite. In all three configurations, realized distance D increases monotonically but less than proportionally with ϵ, ordering checkpoints identically to ϵ. This rules out smaller realized perturbations as an explanation for the heterogeneous robustness response. Meanwhile, utilization ratio $\rho = D/\epsilon$ decreases monotonically in all three configurations (Table 4), making the configured radius a progressively weaker description of the perturbation used for training under $K = 10$. A one-seed $K = 100$ Zephyr diagnostic reaches approximately full utilization, showing that reduced utilization under $K = 10$ does not reflect an intrinsically unreachable constraint boundary. Larger realized perturbations are likewise not, in themselves, evidence of a stronger training adversary. D measures the scale of the intervention, not its effectiveness.

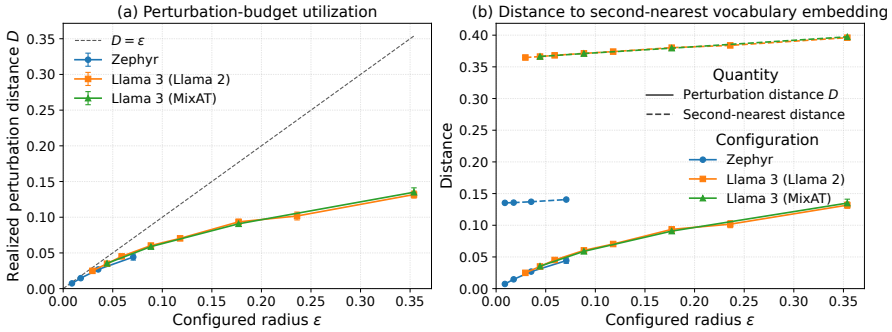

Figure 4: **Larger configured radii produce larger realized perturbations, which remain small relative to the local vocabulary scale.** (a) Realized perturbation distance D as a function of configured radius ϵ; the dashed line indicates $D = \epsilon$. (b) D compared with the distance from the perturbed embedding to its second-nearest vocabulary embedding, which is the nearest alternative when the starting token remains nearest, measured at the perturbed training positions.

5.2 The Scale of the Perturbation Relative to Other Embeddings

CAT perturbations remain local to their original token embeddings over the evaluated range, despite the expanding feasible region. Enlarging the feasible region also raises a geometric question: does the realized perturbation remain local to the original token embedding (Figure 4(b))? At the largest radii, D remains below one quarter of the median nearest-neighbor distance for both models. At the perturbed training positions, the original token remains nearest at every measured Zephyr position and at approximately 99% of Llama 3 positions. By contrast, Soft-Prompt displacements are substantially larger than realized CAT perturbations, despite both operating in the input-embedding space (Appendix C.2). This scale mismatch means that CAT training explores a much smaller neighborhood than the attack, but the CAT sweep does not show that increasing realized training displacement closes the gap. Soft-Prompt robustness does improve as the perturbation budget increases across the 10%-step LAT sweep, although LAT's different perturbation sites and optimization prevent a scale comparison. Moreover, Soft-Prompt displacements vary little while its ASR changes non-monotonically, so magnitude alone does not explain the robustness pattern.

5.3 Between-Seed Variability

Increased between-seed variability occurs at some larger radii, but is not a systematic consequence of increasing the radius. The increase is most apparent for benign compliance: Harmless standard deviation rises from 1.44 at the smallest radius to 26.96 at $\epsilon = 0.08848$ in the Llama 3 (Llama 2 recipe) sweep and reaches 41.61 at the largest MixAT radius. This pattern is neither monotonic nor consistent across evaluations, and three seeds are insufficient to establish that training instability generally increases with radius. We therefore treat variability as a secondary challenge of some larger-radius settings rather than a general effect of increasing ϵ.

6 Can Over-Refusal Supervision Recover Benign Compliance?

Section 4 identifies benign compliance as the most severely affected utility component in CAT. We test whether targeted supervision can recover it by replacing part of CAT's UltraChat utility data with benign OR-Bench prompts (Cui et al., 2024), paired with base-model responses, while holding the adversarial data, perturbation budget, optimization procedure, sampling probabilities, and utility-set size fixed. Under Llama 3 (MixAT), we test replacement sizes of 100–3,000 examples (1.2–35.7% of utility data). The 1,000-example intervention is also evaluated in the other configurations. We use this as a diagnostic rather than proposing it as a defense. Full intervention details are in Appendix D.

The intervention clearly learns the targeted behavior, but this does not consistently transfer to broader benign compliance and comes with a robustness cost. On held-out OR-Bench prompts, the 1,000-example replacement raises success from 46–56% to 93–98% at the first three MixAT radii, with the same qualitative improvement in the other configurations (Figure 5; Appendix D.1).

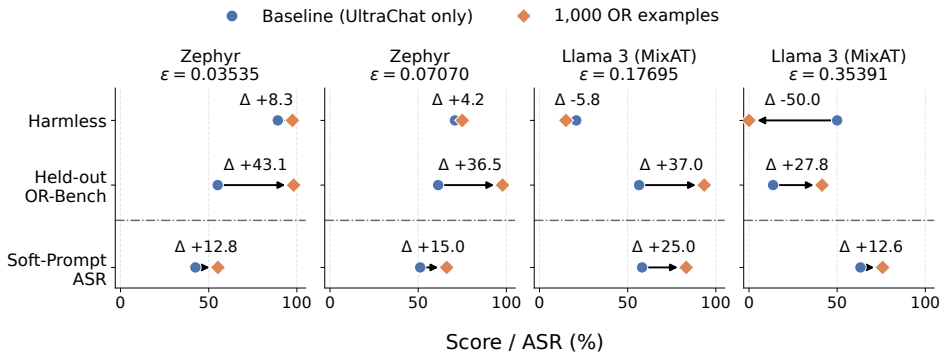

Figure 5: **Targeted over-refusal supervision generalizes within OR-Bench but not consistently to broader benign compliance, while degrading Soft-Prompt robustness.** Change after replacing 1,000 UltraChat examples with OR-Bench for Zephyr and Llama 3 (MixAT recipe). Higher is better for Harmless and held-out OR-Bench success, while lower is better for Soft-Prompt ASR. The complete cross-configuration comparison is in Appendix D.2.

However, this does not consistently transfer to independent benign-compliance evaluations: only one of the sixteen MixAT radius–replacement settings exceeds the corresponding Harmless baseline by more than one baseline standard deviation, while fourteen have lower means. Robustness degrades more consistently. With 1,000 OR examples, Soft-Prompt ASR exceeds the corresponding baseline mean at all twelve configuration–radius combinations, in eleven cases by more than one baseline standard deviation. Figure 5 shows Zephyr and Llama 3 (MixAT recipe). The complete comparison is reported in Appendix D.2. Weakening the targeted supervision does not resolve this: sweeping a separate loss weight for the over-refusal examples at a fixed radius under the MixAT recipe recovers some GCG robustness at the lowest weights but leaves Soft-Prompt ASR above the baseline range at every weight, and Harmless never exceeds the baseline between-seed variation (Appendix D.3).

7 Discussion

Limitations. Our experiments cover CAT and LAT over finite perturbation-budget ranges with three seeds per setting. They therefore identify observed robustness–utility patterns rather than a universal optimal radius. Robustness is conditional on the evaluated attacks, budgets, and judges, and benign-compliance measures sometimes disagree. The perturbation diagnostics are specific to CAT's $K = 10$ inner optimization, while the over-refusal intervention is limited to CAT and primarily one configuration. CAT and LAT are only two latent-space adversarial-training algorithms, but both are established methods on which subsequent work has built. Broader evaluation across methods, models, and training objectives is needed to determine which relationships generalize.

8 Conclusion

We asked whether training with a larger perturbation radius helps in latent-space adversarial training. Across CAT and LAT, larger radii can improve robustness over part of the evaluated range, but the benefit depends on the attack and training configuration. The corresponding utility cost is similarly uneven: benign compliance can deteriorate substantially before capability in CAT, while LAT shows a different utility pattern. Consequently, in CAT, increasing the radius can create meaningful robustness–utility trade-offs, while sufficiently large settings can become dominated when further utility loss is not accompanied by greater measured robustness.

Our CAT analyses further show that this heterogeneous response occurs even though larger configured radii consistently produce larger realized training perturbations. Larger-radius training also raises practical and interpretive challenges, while targeted supervision does not straightforwardly recover CAT's benign-compliance cost in the tested setting. Overall, our results show that perturbation radius is not a direct predictor of the robustness produced by latent-space adversarial training. Its effects must be evaluated across attacks and distinct components of utility, rather than assuming that a larger perturbation budget produces a uniformly stronger robust model.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising text during manuscript preparation. This included suggesting alternative formulations, identifying unclear or redundant passages, improving clarity and concision, and discussing the presentation of results. All AI-assisted text and suggested interpretations were reviewed against the underlying results and revised by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies jailbreak robustness and therefore involves dual-use techniques that could also inform attempts to circumvent model safeguards. We use established attack methods and evaluation datasets to characterize existing defenses, with the goal of improving the evaluation and design of robust language models.

REPRODUCIBILITY STATEMENT

We provide training configurations and perturbation-budget sweeps (Appendix A.1, Tables 1 and 2), evaluation protocols (Appendix A.2, Table 3), complete numerical results (Appendix B), perturbation diagnostics (Appendix C), and intervention details (Appendix D). All base models, training datasets, and attacks are publicly available. We report training seeds and distinguish three-seed results from single-seed diagnostic and intervention experiments where applicable.

REFERENCES

Tim Beyer, Sophie Xhonneux, Simon Geisler, Gauthier Gidel, Leo Schwinn, and Stephan Günnemann. Position: Llm-safety evaluations lack robustness. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Csaba Dékány, Stefan Balauca, Dimitar I Dimitrov, Robin Staab, and Martin Vechev. Mixat: Combining continuous and discrete adversarial training for llms. *Advances in Neural Information Processing Systems*, 38:86727–86760, 2026.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, 2023.

Shaopeng Fu and Di Wang. Understanding and improving continuous adversarial training for llms via in-context learning theory. *arXiv preprint arXiv:2604.12817*, 2026.

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.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Lin Li and Michael Spratling. Data augmentation alone can improve adversarial training. In *International Conference on Learning Representations*, 2023.

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*, volume 2024, pp. 56174–56194, 2024.

David Lüdke, Tom Wollschläger, Paul Ungermann, Stephan Günnemann, and Leo Schwinn. Diffusion llms are natural adversaries for any llm. *arXiv preprint arXiv:2511.00203*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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.

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)*, pp. 5377–5400, 2024.

Leo Schwinn and Simon Geisler. Revisiting the robust alignment of circuit breakers. *arXiv preprint arXiv:2407.15902*, 2024.

Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Günnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *Advances in Neural Information Processing Systems*, 37:9086–9116, 2024.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, et al. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

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.

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.

A Experimental Details

This section provides the additional configuration details needed to reproduce the controlled perturbation-radius sweeps and evaluations. The CAT experiments consist of 16 radius settings across three model–recipe configurations, with three independently trained checkpoints per setting (48 sweep checkpoints in total). The LAT experiments consist of two four-setting sweeps with three seeds per setting (24 LAT sweep checkpoints). The unmodified $\epsilon = 0$ models serve only as reference conditions.

A.1 Training Configurations and Perturbation Sweeps

CAT configuration. All configurations use the released CAT training implementation, with the perturbation-radius sweep and instrumentation described below. The adversarial objective weights are $\alpha_t = 0.5$, $\alpha_a = 0.5$, and $\alpha_u = 1.0$. Training uses LoRA on all linear layers and the shared optimization settings in Table 1. Within each model–recipe configuration, only the perturbation radius is varied.

Table 1: **CAT training configuration and controlled perturbation-radius sweeps.** All sweep settings use three training seeds (42, 123, and 456). Parameters not listed as recipe-specific are shared across the three configurations.

Shared training settings	
LoRA rank / alpha / dropout	64 / 16 / 0.1
LoRA target modules	All linear layers
Quantization	4-bit NF4
Outer optimizer	Paged AdamW, 32-bit
Learning rate	2×10^{-4}
LR schedule	Cosine, warmup ratio 0.03
Weight decay	0.001
Effective batch size	64
Maximum sequence length	256
Inner iterations	$K = 10$
Inner optimizer	Signed gradient descent
Initial perturbation	$\delta_0 = 0$
Projection	After every inner step
Perturbed positions	Non-padding positions outside response
Adversarial / utility sampling	0.125 / 0.875
Toward / away / utility weights	0.5 / 0.5 / 1.0
Sweep update norm	0.1ϵ before projection
Training seeds	42, 123, 456

Configuration	**Epochs**	**Away cutoff**	(s, ϵ) **sweep**
Zephyr	5	-5.0	$(0.05, 0.00884)$, $(0.10, 0.01768)$, $(0.20, 0.03535)$, $(0.40, 0.07070)$
Llama 3 (Llama 2 recipe)	2	-7.5	$(0.05, 0.02949)$, $(0.075, 0.04424)$, $(0.10, 0.05898)$, $(0.15, 0.08848)$, $(0.20, 0.11797)$, $(0.30, 0.17695)$, $(0.40, 0.23594)$, $(0.60, 0.35391)$
Llama 3 (MixAT recipe)	2	-5.0	$(0.075, 0.04424)$, $(0.15, 0.08848)$, $(0.30, 0.17695)$, $(0.60, 0.35391)$

For the controlled sweeps, the coordinate-wise signed-gradient step is

$$\eta = \frac{0.1\epsilon}{\sqrt{d}}, \tag{4}$$

so that a dense, unprojected signed-gradient update has ℓ_2 norm 0.1ϵ. This keeps the update magnitude fixed relative to the configured radius as ϵ changes. The released implementation's radius-relative option does not include this dimension normalization. We use the form above specifically to make the radius sweep controlled, not as a claim that it is a superior CAT optimization procedure.

At each training step, we measure realized perturbation distance D from the lowest-loss inner-loop iterate selected for adversarial training. We average $\|\delta_i^*\|_2$ over perturbed positions and training steps, and define utilization as $\rho = D/\epsilon$.

LAT configuration. LAT uses `Llama 3-8B-Instruct` and perturbs the embedding output and the MLP outputs of decoder layers 8, 16, 24, and 30. A separate perturbation is maintained at each location and constrained independently for each token to an ℓ_2 ball of radius ϵ, with projection after every inner update. Perturbations are initialized at zero and optimized for $K = 16$ iterations using AdamW, with new optimizer state for each attack. The final iterate is used for training. The released configuration uses $\epsilon = 6$ and inner learning rate 10^{-3}. Our LAT sweeps vary only ϵ and the inner learning rate, with all remaining training settings fixed as reported in Table 2.

For a freshly initialized AdamW optimizer and zero perturbation, the first update is approximately sign-normalized, giving a per-token update norm $\|\Delta_1\|_2 \approx \mathrm{LR}\sqrt{4096} = 64\,\mathrm{LR}$. We therefore set $\mathrm{LR} = \epsilon/640$ in the 10%-step sweep so that the first update is approximately 0.1ϵ; this relation applies only to the first AdamW iteration.

Table 2: **LAT training configuration and controlled perturbation-radius sweeps.** All settings use three training seeds (42, 123, and 456). Across the sweeps, only ϵ and the inner learning rate are varied.

Shared training settings	
Base model	`Llama 3-8B-Instruct`
LoRA rank / alpha / dropout	64 / 8 / 0
LoRA target modules	q, k, v, o, up, down projections
Quantization	None
Outer optimizer	AdamW, LR 8×10^{-5}
Outer weight decay	0.01
Outer steps	100 (4 model updates per step)
Maximum sequence length	2048
Inner iterations	$K = 16$
Inner optimizer	AdamW
Inner optimizer β / eps / weight decay	$(0.9, 0.999)$ / 10^{-8} / 0.01
Initial perturbation	Zero
Projection	Per-token, per-location ℓ_2 ball after each step
Perturbed locations	Embedding; MLP outputs of layers 8, 16, 24, 30
Training seeds	42, 123, 456

Scaling	ϵ	**Inner LR**
Original-LR	6	0.001
	12	0.002
	24	0.004
	48	0.008
10%-step	6	0.009375
	12	0.01875
	24	0.0375
	48	0.075

Table 3: **Robustness and utility evaluation protocols.**

Evaluation	Implementation source	Set / N	Configuration	Judge / scoring
GCG	(Dékány et al., 2026)	HarmBench subset / 40	500 steps; 512 candidates/step	HarmBench classifier
AutoDAN	(Dékány et al., 2026)	HarmBench subset / 40	64 population; 100 generations	HarmBench classifier
Soft-Prompts	(Schwinn & Geisler, 2024)	HarmBench test / 159	20 soft vectors; $\geq$ 500 steps	HarmBench classifier
Inpainting	(Lüdke et al., 2025)	JailbreakBench / 100	Best-of-1,024 pre-generated variants	StrongREJECT > 0.5
MMLU	(Dékány et al., 2026)	Standard / –	Multiple-choice likelihood	Accuracy; no chat template
ARC-E / C	(Dékány et al., 2026)	Standard / –	Multiple-choice likelihood	Accuracy; no chat template
Harmless	(Dékány et al., 2026)	40	Greedy generation	Successful-response judge
XSTest	(Röttger et al., 2024)	250 safe / 200 unsafe	Greedy generation	16-prefix refusal classifier

A.2 Evaluation Protocols

Table 3 summarizes the attack configurations. Attack success rates are comparable only within an attack family: the attacks use different datasets, budgets, generation procedures, and judges. Each attack pipeline uses a fixed evaluation seed independently of the checkpoint's training seed, so variation reported across the three training seeds reflects differences between independently trained checkpoints rather than independently sampled attack randomness. GCG and AutoDAN generate up to 512 tokens per completion. Soft-Prompts, Inpainting, and XSTest generate up to 256 tokens.

For Harmless, we report the proportion of prompts judged to receive a successful response. We apply a rule-based correction that assigns a score of zero to completions consisting of the fixed CAT refusal continuation, even when the LLM judge labels them as successful.

For XSTest, safe compliance is the fraction of 250 safe prompts that are not classified as refusals, while unsafe refusal is the fraction of 200 unsafe prompts that are classified as refusals. The classifier labels a completion as a refusal when its normalized text begins with one of 16 predefined refusal phrases. A post-hoc comparison with an auxiliary LLM judge revealed checkpoint-dependent disagreement, so we interpret small XSTest differences cautiously. The auxiliary judge is not treated as ground truth or used to modify the reported scores.

B Complete Sweep Results

This section reports the complete numerical results for the controlled CAT and LAT sweeps. Values are mean $\pm$ sample standard deviation over the three independently trained checkpoints at each radius. Unmodified and released-model reference values are reported without across-training-seed standard deviations.

B.1 Realized Perturbations

Table 4 reports configured radius ϵ, realized perturbation distance D, and utilization $\rho = D/\epsilon$ for every sweep setting. Within each of the three configurations, D increases monotonically with ϵ and ρ decreases monotonically. ϵ is the configured experimental variable, D is the realized perturbation distance, and ρ is a diagnostic of how much of the feasible region the finite inner optimization uses. None of these quantities is a measure of adversarial-training effectiveness.

B.2 Utility

Tables 5 and 6 report the complete CAT and LAT capability and behavioral utility results, respectively.

The relatively large standard deviations at several behavioral settings motivate caution against interpreting fine-grained differences between radius settings. In particular, the apparent increase in mean Harmless at the largest MixAT radius should not be interpreted as evidence of a stable recovery in benign compliance.

Tables 5 and 7 also report the standard deviations underlying the CAT variability in Section 5.3. The increased variability at some intermediate and larger radii is most apparent for benign compliance, but is neither monotonic nor consistent across all evaluations. These values describe variation in

Table 4: **Configured and realized perturbation magnitude.** D is the mean realized perturbation distance of the lowest-loss inner-loop iterate selected for training, and $\rho = D/\epsilon$.

Configuration	s	ϵ	D	ρ
Zephyr	0.05	0.00884	0.00734 ± 0.00003	0.831
	0.10	0.01768	0.01466 ± 0.00006	0.829
	0.20	0.03535	0.02688 ± 0.00068	0.760
	0.40	0.07070	0.04422 ± 0.00423	0.625
Llama 3 (Llama 2)	0.05	0.02949	0.02492 ± 0.00050	0.845
	0.075	0.04424	0.03491 ± 0.00107	0.789
	0.10	0.05898	0.04517 ± 0.00030	0.766
	0.15	0.08848	0.06035 ± 0.00039	0.682
	0.20	0.11797	0.07039 ± 0.00285	0.597
	0.30	0.17695	0.09350 ± 0.00361	0.528
	0.40	0.23594	0.10172 ± 0.00553	0.431
	0.60	0.35391	0.13176 ± 0.00496	0.372
Llama 3 (MixAT)	0.075	0.04424	0.03539 ± 0.00045	0.800
	0.15	0.08848	0.05884 ± 0.00232	0.665
	0.30	0.17695	0.09078 ± 0.00334	0.513
	0.60	0.35391	0.13500 ± 0.00616	0.381

Table 5: **Complete utility results for the CAT sweeps.** Positive-radius values are mean $\pm$ sample standard deviation over three independently trained checkpoints. The $s = 0$ rows report the unmodified base models, plotted at $\epsilon = 0$ as reference conditions, and therefore have no across-training-seed standard deviation. Higher is better. SC and UR denote XSTest safe compliance and unsafe refusal, respectively.

Configuration	ϵ	MMLU	ARC-E	ARC-C	Harmless	XSTest SC	XSTest UR
Zephyr	0	56.02	81.00	55.20	100.00	94.80	9.50
	0.00884	55.05 ± 0.30	79.00 ± 0.26	51.17 ± 0.35	98.33 ± 1.44	62.40 ± 7.34	98.83 ± 0.58
	0.01768	54.98 ± 0.26	78.52 ± 0.36	51.08 ± 0.50	94.17 ± 3.82	54.67 ± 6.47	98.83 ± 0.76
	0.03535	55.17 ± 0.17	78.04 ± 0.40	50.94 ± 0.76	89.17 ± 1.44	53.47 ± 2.95	99.00 ± 0.50
	0.07070	54.92 ± 0.17	77.08 ± 0.88	49.32 ± 0.82	70.83 ± 9.46	70.93 ± 13.05	89.00 ± 9.37
Llama 3 (Llama 2)	0	60.75	79.12	49.40	100.00	97.20	91.50
	0.02949	60.10 ± 0.09	80.95 ± 0.30	51.77 ± 0.43	95.83 ± 1.44	61.87 ± 5.72	99.33 ± 0.29
	0.04424	59.74 ± 0.17	80.80 ± 0.44	51.45 ± 0.38	90.83 ± 7.22	48.13 ± 18.36	92.67 ± 10.97
	0.05898	59.83 ± 0.13	80.78 ± 0.18	51.31 ± 0.39	76.67 ± 7.64	47.47 ± 14.14	99.67 ± 0.29
	0.08848	59.35 ± 0.55	79.45 ± 0.81	49.69 ± 1.00	45.83 ± 26.96	52.13 ± 29.31	96.83 ± 3.51
	0.11797	58.81 ± 0.11	78.05 ± 1.21	49.72 ± 0.35	49.17 ± 20.05	32.40 ± 26.89	99.17 ± 0.29
	0.17695	57.69 ± 0.45	74.99 ± 1.58	47.32 ± 0.40	7.50 ± 12.99	20.67 ± 22.10	98.50 ± 0.87
	0.23594	56.23 ± 1.04	73.18 ± 2.82	46.22 ± 1.92	2.50 ± 4.33	32.00 ± 28.93	98.83 ± 1.26
	0.35391	54.94 ± 0.98	47.97 ± 4.30	37.37 ± 2.10	1.67 ± 2.89	0.80 ± 0.40	100.00 ± 0.00
Llama 3 (MixAT)	0	60.75	79.12	49.40	100.00	97.20	91.50
	0.04424	59.73 ± 0.22	80.46 ± 0.43	50.85 ± 0.39	82.50 ± 13.92	51.60 ± 15.48	99.33 ± 0.29
	0.08848	59.47 ± 0.14	79.44 ± 0.29	49.34 ± 0.65	58.33 ± 14.65	51.07 ± 10.51	98.50 ± 0.87
	0.17695	58.51 ± 0.68	75.45 ± 2.54	47.84 ± 1.65	20.83 ± 15.07	60.27 ± 10.01	97.00 ± 1.80
	0.35391	55.74 ± 0.41	41.29 ± 19.91	32.37 ± 9.46	50.00 ± 41.61	37.33 ± 10.99	99.50 ± 0.50

downstream outcomes across three independently trained checkpoints and are not interpreted as estimates of training instability.

B.3 Adversarial Robustness

Tables 7 and 8 report the complete CAT and LAT adversarial-robustness results, respectively. Inpainting was evaluated only for CAT.

B.4 Observed Non-Dominated Settings

Table 9 reports the observed non-dominated settings for each robustness–utility pairing. If the unmodified $s = 0$ reference models are additionally included in the dominance comparison, they are

Table 6: **Complete utility results for the LAT sweeps.** Positive-radius values are mean ± sample standard deviation over three independently trained checkpoints. The unmodified Llama 3 and released LAT checkpoints are reference conditions and therefore have no across-training-seed standard deviation. Higher is better. SC and UR denote XSTest safe compliance and unsafe refusal, respectively.

Configuration	ϵ	MMLU	ARC-E	ARC-C	Harmless	XSTest SC	XSTest UR
Unmodified Llama 3	0	60.75	79.12	49.40	100.00	97.20	91.50
Released LAT	6	58.23	73.19	42.66	97.50	86.00	7.00
Original-LR scaling	6	48.05 ± 2.25	48.87 ± 0.61	28.10 ± 0.44	91.67 ± 2.89	96.13 ± 0.46	4.33 ± 0.58
	12	54.72 ± 5.55	52.16 ± 18.83	32.88 ± 9.92	97.50 ± 0.00	96.80 ± 0.69	4.50 ± 0.87
	24	59.59 ± 0.53	57.45 ± 6.01	37.09 ± 2.18	100.00 ± 0.00	99.20 ± 0.40	3.67 ± 0.29
	48	60.15 ± 0.09	78.33 ± 1.36	48.04 ± 0.72	100.00 ± 0.00	99.47 ± 0.23	3.50 ± 0.00
10%-step scaling	6	60.34 ± 0.13	78.88 ± 0.49	47.81 ± 0.27	100.00 ± 0.00	99.33 ± 0.23	3.33 ± 0.29
	12	60.21 ± 0.15	79.10 ± 0.11	48.78 ± 0.39	100.00 ± 0.00	99.20 ± 0.00	3.33 ± 0.29
	24	59.61 ± 1.60	75.44 ± 6.21	47.52 ± 3.63	100.00 ± 0.00	99.07 ± 0.23	2.83 ± 0.76
	48	60.68 ± 0.14	79.22 ± 0.14	49.29 ± 0.22	100.00 ± 0.00	99.20 ± 0.00	3.17 ± 0.29

Table 7: **Complete adversarial-robustness results for the CAT sweeps.** Values are mean ASR ± sample standard deviation over three independently trained checkpoints. The $s = 0$ rows report the unmodified base models, plotted at $\epsilon = 0$ as reference conditions, and therefore have no across-training-seed standard deviation. Lower is better. ASRs are comparable within attack columns, not across attacks. "–" indicates an unavailable Inpainting evaluation.

Configuration	ϵ	GCG	AutoDAN	Soft-Prompts	Inpainting
Zephyr	0	92.50	90.00	74.09	100.00
	0.00884	19.17 ± 3.82	1.67 ± 1.44	43.44 ± 4.53	91.67 ± 4.93
	0.01768	17.50 ± 6.61	2.50 ± 0.00	43.40 ± 5.42	95.67 ± 2.52
	0.03535	18.33 ± 10.10	0.83 ± 1.44	42.45 ± 13.65	95.67 ± 1.15
	0.07070	9.17 ± 1.44	0.00 ± 0.00	51.13 ± 3.21	96.00 ± 3.46
Llama 3 (Llama 2)	0	28.33	10.00	75.41	97.00
	0.02949	22.50 ± 6.61	1.67 ± 2.89	52.62 ± 2.55	94.00 ± 5.29
	0.04424	20.83 ± 3.82	0.00 ± 0.00	57.29 ± 9.37	–
	0.05898	16.67 ± 10.41	0.00 ± 0.00	48.32 ± 1.89	89.67 ± 3.06
	0.08848	22.50 ± 12.99	0.00 ± 0.00	59.25 ± 16.08	–
	0.11797	19.17 ± 7.22	0.00 ± 0.00	51.26 ± 3.17	88.67 ± 2.52
	0.17695	19.17 ± 11.55	0.00 ± 0.00	59.33 ± 1.41	–
	0.23594	22.50 ± 5.00	0.00 ± 0.00	62.53 ± 5.64	93.67 ± 3.21
	0.35391	33.33 ± 9.46	0.00 ± 0.00	64.46 ± 2.30	–
Llama 3 (MixAT)	0	28.33	10.00	75.41	97.00
	0.04424	25.83 ± 1.44	0.83 ± 1.44	55.47 ± 4.46	87.67 ± 6.51
	0.08848	12.50 ± 2.50	0.00 ± 0.00	43.23 ± 3.92	89.67 ± 4.04
	0.17695	17.50 ± 6.61	0.00 ± 0.00	58.18 ± 4.86	94.00 ± 1.73
	0.35391	25.00 ± 10.90	0.00 ± 0.00	63.21 ± 3.20	89.67 ± 1.15

Four Inpainting evaluations in the Llama 3 (Llama 2 recipe) sweep are unavailable because the pre-generated attack dataset became unavailable before those evaluations were completed. We did not substitute a replacement dataset, as doing so would have changed the evaluation condition.

Table 8: **Complete adversarial-robustness results for the LAT sweeps.** Values are mean ASR $\pm$ sample standard deviation over independently trained checkpoints. The unmodified Llama 3 and released LAT checkpoints are reference conditions and therefore have no across-training-seed standard deviation. Lower is better, and ASRs are compared only within attack families.

Configuration	ϵ	GCG	AutoDAN	Soft-Prompts
Unmodified Llama 3	0	28.33	10.00	75.41
Released LAT	6	7.50	0.00	62.58
Original-LR scaling	6	7.50 ± 6.61	0.00 ± 0.00	38.15 ± 7.79
	12	15.83 ± 6.29	0.00 ± 0.00	82.79 ± 1.16
	24	21.67 ± 3.82	2.50 ± 4.33	84.76 ± 1.26
	48	12.50 ± 5.00	1.67 ± 1.44	76.67 ± 1.51
10%-step scaling	6	16.67 ± 1.44	5.00 ± 4.33	79.98 ± 1.68
	12	17.50 ± 5.00	5.00 ± 0.00	63.92 ± 8.91
	24	15.83 ± 17.01	6.67 ± 6.29	52.17 ± 5.70
	48	21.67 ± 3.82	6.67 ± 1.44	50.44 ± 7.60

Table 9: **Observed non-dominated radius settings.** Entries give configured radii ϵ for which no other evaluated radius in the same model–recipe sweep has both at least as much utility and no greater ASR, with at least one strict improvement. Dominance is computed from three-seed means. These are observed non-dominated settings over a finite sweep, not estimates of a continuous Pareto frontier. SP denotes Soft-Prompts.

Configuration	Harmless–GCG	Harmless–SP	MMLU–GCG	MMLU–SP
Zephyr	0.00884, 0.01768, 0.07070	0.00884, 0.01768, 0.03535	0.01768, 0.03535, 0.07070	0.03535
Llama 3 (Llama 2)	0.02949, 0.04424, 0.05898	0.02949, 0.05898	0.02949, 0.05898	0.02949, 0.05898
Llama 3 (MixAT)	0.04424, 0.08848	0.04424, 0.08848	0.04424, 0.08848	0.04424, 0.08848

	ARC-E		ARC-C		XSTest SC	
Configuration	GCG	SP	GCG	SP	GCG	SP
Zephyr	.00884, .01768, .07070	.00884, .01768, .03535	.00884, .01768, .07070	.00884, .01768, .03535	.07070	.00884, .01768, .03535, .07070
Llama 3 (Llama 2)	.02949, .04424, .05898	.02949, .05898	.02949, .04424, .05898	.02949, .05898	.02949, .04424, .05898	.02949, .05898
Llama 3 (MixAT)	.04424, .08848	.04424, .08848	.04424, .08848	.04424, .08848	.08848, .17695	.04424, .08848, .17695

non-dominated in almost all robustness–utility pairings. We exclude them from Table 9 because the main analysis compares positive-radius CAT settings within the controlled radius sweeps.

The exact set of non-dominated settings should be interpreted cautiously when neighboring means are close or between-seed variation is large. In particular, dominance between two three-seed means does not imply that every independently trained checkpoint at one radius dominates every checkpoint at another. The XSTest exceptions occur at settings where XSTest and Harmless disagree and where classifier limitations and between-seed variation make the fine-grained ordering less reliable.

C Perturbation Diagnostics

This section reports the measurements underlying Section 5 and two diagnostics examining alternative explanations for the relationship between configured and realized perturbations. None of these quantities is used as an additional measure of robustness.

C.1 Locality Relative to Vocabulary Embeddings

Two reference scales are used to interpret the realized perturbation distance D. First, Table 10 reports the distribution of ℓ_2 distances from each token embedding to its nearest neighbor in the frozen vocabulary embedding matrix. The median distance is 0.179 for Zephyr and 0.596 for Llama 3. At the largest radii, realized D remains below one quarter of the corresponding median in both model families.

We additionally measure locality directly at the perturbed training positions. The original token remains the closest vocabulary embedding at every measured Zephyr position across the sweep. For

Table 10: **Distribution of ℓ_2 distances from each vocabulary embedding to its nearest neighbor.** Computed on each model's frozen embedding matrix. Percentiles are reported instead of the minimum, because a small number of untrained vocabulary entries have almost identical embeddings, producing minima close to zero.

Model	$\lvert V\rvert$	p1	p25	p50	p75	p99
Zephyr	32,000	0.109	0.167	0.179	0.188	0.207
Llama 3	128,256	0.288	0.553	0.596	0.626	0.711

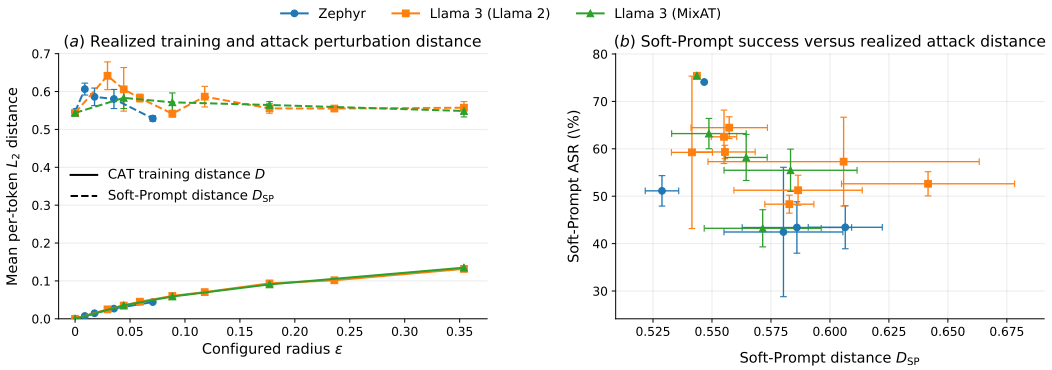

Figure 6: **Soft-Prompt displacement does not explain the robustness response to CAT radius.** (a) Realized CAT training perturbation distance D (solid lines) and Soft-Prompt displacement D_{SP} (dashed lines) across the configured CAT radii. Both quantities are measured in the input-embedding coordinate system but describe different interventions and are compared only in scale. Positive-radius points show means across three independently trained checkpoints, with error bars indicating one sample standard deviation. $\epsilon = 0$ denotes the unmodified reference model and therefore has no across-training-seed standard deviation. (b) Soft-Prompt ASR against D_{SP} for the same checkpoints, with horizontal and vertical error bars indicating one sample standard deviation across training seeds for positive-radius settings. Lower ASR is better. Soft-Prompt displacement varies comparatively little across the sweeps and does not provide a consistent ordering of Soft-Prompt robustness.

Llama 3, it remains closest at approximately 99% of positions even at the largest radius. The distance to the nearest alternative token also remains substantially larger than the realized perturbation. Thus, over the evaluated $K = 10$ sweeps, increasing the configured radius moves embeddings further from their starting tokens without generally making another vocabulary embedding the nearest one.

The configured and realized regions should nevertheless be distinguished. At the largest Llama 3 setting, $\epsilon = 0.35391$ is comparable to the local distance to the nearest alternative vocabulary entry, even though the realized perturbation remains substantially smaller. These measurements characterize the perturbations produced by the evaluated inner optimization and do not establish that another vocabulary embedding could not become nearest under a different optimization procedure or budget.

C.2 Soft-Prompt Displacement

Figure 6 compares the displacement realized during CAT training with that realized by the Soft-Prompt attack. Let D_{SP} denote the mean per-soft-token ℓ_2 displacement from initialization. Both D and D_{SP} are measured in the model's input-embedding coordinate system, but describe different interventions and are therefore compared only in scale.

Across the controlled sweeps, D_{SP} is substantially larger than D and varies comparatively little, while Soft-Prompt ASR does not order consistently with D_{SP} (Figure 6). This rules out a simple displacement-magnitude explanation for the Soft-Prompt robustness pattern without identifying the mechanism responsible for it.

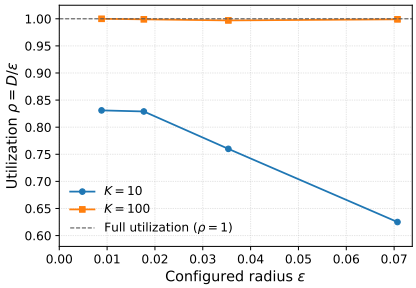

Figure 7: **Increasing the inner-loop budget makes the constraint boundary reachable in the Zephyr diagnostic.** Utilization $\rho = D/\epsilon$ under the $K = 10$ training procedure used in the controlled sweeps and a one-seed $K = 100$ diagnostic. The dashed line denotes full utilization, $\rho = 1$. The $K = 10$ points are means over three training seeds, whereas the $K = 100$ diagnostic uses one seed.

C.3 Increasing the Inner-Loop Budget

The main sweeps use $K = 10$ inner iterations, with utilization $\rho = D/\epsilon$ decreasing as ϵ increases. To test whether this means that the constraint boundary is intrinsically unreachable, we repeat the four-point Zephyr sweep with $K = 100$ while keeping the radius-relative update coefficient fixed at $\lambda = 0.1$. The diagnostic uses one training seed.

With $K = 100$, utilization is 1.000, 0.999, 0.997, and 0.999 across the four radii, compared with 0.831, 0.829, 0.760, and 0.625 under the $K = 10$ sweep. Thus, declining utilization under $K = 10$ does not imply that the constraint boundary is intrinsically unreachable, since for Zephyr it becomes reachable when the inner optimization budget is increased. This diagnostic does not show that fuller utilization produces a better adversarial objective or a more robust trained model. The $K = 100$ checkpoints were not evaluated for robustness or utility, and the diagnostic was not repeated for Llama 3.

C.4 Selected versus Final Inner-Loop Iterate

CAT trains on the lowest-loss inner-loop iterate, which is not necessarily the final iterate, so we test whether this selection explains the declining utilization. It does not. Across every sweep setting, the median selected iteration is 10 and the mean is always above 9.2. Recomputing D from the final $K = 10$ iterate gives values close to the loss-selected distances and preserves the same less-than-proportional relationship between D and ϵ. The vocabulary-locality measurements are likewise materially unchanged. Thus, the configured-versus-realized perturbation and locality results are not explained by whether the selected or final inner-loop iterate is measured.

D Targeted Over-Refusal Supervision

This section reports the setup and complete available results for the intervention in Section 6.

CAT's utility objective uses a small, fixed prefix of UltraChat. The intervention replaces the first n UltraChat conversations of that prefix with n benign OR-Bench prompts (Cui et al., 2024), keeping the adversarial data, perturbation radius, optimization procedure, interleaving procedure, sampling probabilities, and total utility count unchanged. OR-Bench contains $80{,}359$ benign prompts that were generated by rewriting harmful seeds and retained when a majority of an ensemble of judges labeled them safe. The prompts are split deterministically and without overlap into the n examples used for training and the remaining $80{,}359 - n$ used to evaluate generalization within the same distribution.

OR-Bench provides prompts but no target responses. Each training prompt is paired with a response generated by greedy decoding from the corresponding base model before adversarial training, so Zephyr training uses Zephyr-generated responses and Llama 3 training uses Llama 3-generated re-

Table 11: **Held-out OR-Bench success, Llama 3 (MixAT recipe).** Success rate on the OR-Bench prompts not used for training (out of $80{,}359 - n$ for replacement size n), comparing checkpoints trained on UltraChat alone against checkpoints with n UltraChat examples replaced by OR examples. Percentages give the replaced OR examples as a fraction of the total utility data. Values with $\pm$ are mean $\pm$ sample standard deviation over three independently trained checkpoints. Values without $\pm$ come from a single training run.

s	UltraChat only	$n = 100$ (1.2%)	$n = 500$ (6.0%)	$n = 1{,}000$ (11.9%)	$n = 3{,}000$ (35.7%)
0.075	55.84 ± 7.53	91.86	96.63	98.40	87.36
0.150	46.20 ± 6.03	94.99	95.78 ± 1.56	96.50	87.04
0.300	56.47 ± 8.42	77.74	95.50	93.42	86.06
0.600	13.65 ± 12.61	44.28 ± 29.14	54.66 ± 17.28	41.44 ± 37.00	29.69 ± 26.46

sponses. Other target constructions were not tested, so the targeted-learning result below is specific to this response construction.

The main experiment uses the Llama 3 (MixAT recipe) configuration at all four of its sweep radii, with replacement sizes $n = 100$, 500, 1,000, and 3,000, corresponding to approximately 1.2%, 6.0%, 11.9%, and 35.7% of the utility data. This configuration was chosen because benign compliance declines substantially across its sweep without reaching the near-zero values seen at the largest Llama 2-recipe radii, leaving a range in which the intervention could move the measurement in either direction. The $n = 1{,}000$ replacement is additionally evaluated in the other two configurations at their own sweep radii, giving twelve configuration–radius combinations in total.

Most replacement conditions were trained with a single seed. Three-seed results are available only for isolated settings, which are marked with $\pm$ in Table 11. Entries without $\pm$ are single-seed. The unmodified sweep baselines used for comparison throughout this section are the three-seed results already reported in Appendix B.

The loss-weight experiment (Appendix D.3) splits the single utility term into an UltraChat term with weight α_u and an OR term with weight α_{OR}. The base weight $\alpha_{\mathrm{OR}} = 0.0595$ is the proportion of OR examples in the utility data ($500/8{,}400$). α_u is set once at this reference point to $1 - \alpha_{\mathrm{OR}} = 0.9405$ so that introducing the separate term does not also change the overall strength of the utility objective relative to the adversarial terms, and is then kept fixed while α_{OR} is varied. The remaining four weights scale the base value by factors of four in either direction. The cutoff values and $\alpha_t = \alpha_a = 0.5$ are unchanged from the main sweeps.

D.1 Targeted OR-Bench Generalization

Table 11 reports success on the held-out OR-Bench prompts. Every replacement size improves over the corresponding UltraChat-only baseline at every Llama 3 (MixAT) radius tested, by 32–43 percentage points at the smallest radius and 16–41 at the largest. For the $n = 1{,}000$ replacement at the first three radii, success rises from 55.84, 46.20, and 56.47 to 98.4, 96.50, and 93.42. The $n = 1{,}000$ replacement produces the same qualitative improvement in the other two configurations, where success rises from 36–67% without OR training to 82–98% with OR examples at every radius.

D.2 Benign Compliance, Robustness, and Capability under Replacement

Figure 8 reports Harmless, GCG, and Soft-Prompt results across replacement sizes under the Llama 3 (MixAT recipe), against the corresponding three-seed baselines from Appendix B.

Benign compliance. Across the sixteen Llama 3 (MixAT) radius–replacement settings, only one exceeds the corresponding Harmless baseline mean by more than one baseline standard deviation: at $s = 0.075$ with $n = 1{,}000$, Harmless reaches 97.5 against a baseline of 82.50 ± 13.92. Fourteen settings have a lower mean than the baseline and two a higher one, and replacement size shows no consistent ordering. Because most replacement conditions were trained once, this does not establish systematic degradation relative to the unmodified sweep. It provides no evidence that OR replacement generally improves Harmless. XSTest safe compliance is similarly inconsistent and can move

Table 12: **OR loss-weight sweep, Llama 3 (MixAT recipe),** $s = 0.15$**.** Complete results across all five tested weights. The baseline column and the three smaller weights are mean ± sample standard deviation over three independently trained checkpoints. The two largest weights are single-seed. Higher is better for Harmless, MMLU, ARC-E, and ARC-C. Lower is better for the ASR rows.

Metric	Baseline	0.00371875	0.014875	0.0595	0.238	0.952
Harmless	58.33 ± 14.65	67.50 ± 9.01	55.83 ± 13.77	63.33 ± 6.29	50.00	20.00
MMLU	59.47 ± 0.14	59.34 ± 0.26	59.28 ± 0.55	59.41 ± 0.53	59.08	59.60
ARC-E	79.44 ± 0.29	80.32 ± 0.45	79.55 ± 1.57	79.91 ± 0.17	79.08	79.21
ARC-C	49.34 ± 0.65	50.25 ± 0.56	49.63 ± 1.26	49.63 ± 0.56	49.40	48.98
GCG ASR	12.50 ± 2.50	27.50 ± 5.00	15.83 ± 5.20	40.83 ± 6.29	42.50	35.00
AutoDAN ASR	0.00 ± 0.00	0.83 ± 1.44	0.00 ± 0.00	0.00 ± 0.00	0.00	0.00
Soft-Prompt ASR	43.23 ± 3.92	62.41 ± 2.77	60.57 ± 5.35	63.59 ± 3.15	66.86	71.95
XSTest SC	51.07 ± 10.51	41.20 ± 37.59	72.67 ± 11.68	64.27 ± 7.42	60.40	61.60
XSTest UR	98.50 ± 0.87	96.33 ± 5.06	89.50 ± 6.87	95.00 ± 2.29	94.00	97.50

independently of Harmless, and given the checkpoint-dependent limitations of the prefix detector (Appendix A.2) we do not use it to rank the replacement conditions.

Robustness. GCG and Soft-Prompt ASR rise above the corresponding baseline range across the replacement sizes and radii examined. At $s = 0.15$, every Soft-Prompt replacement value is well above the baseline range. AutoDAN remains at or near its floor and does not distinguish the replacement conditions, and Inpainting is unavailable for these experiments, so the robustness conclusion rests on GCG and Soft-Prompts.

Capability. MMLU, ARC-Easy, and ARC-Challenge remain close to the UltraChat-only results. Across all four replacement sizes, every benchmark stays within, or close to, the baseline's own seed-to-seed variation at every radius. MMLU is a partial exception, lying slightly above the baseline range at several replacement sizes and radii, but by no more than two points. OR replacement also does not remove the capability degradation observed at the largest perturbation budget: the decline in ARC performance at $s = 0.6$ is still present.

Cross-configuration. For the $n = 1{,}000$ replacement evaluated in all three configurations (Figure 9), held-out OR-Bench success improves at every configuration–radius combination, confirming that the targeted behavior is learned across configurations. This improvement does not consistently transfer to Harmless: only three combinations exceed the baseline by more than one standard deviation, and the direction of change differs between configurations. Meanwhile, Soft-Prompt ASR exceeds the corresponding baseline mean at all twelve configuration–radius combinations, and by more than one baseline standard deviation in eleven.

D.3 Loss-Weight Sweep

Figure 10 shows the effect of varying the OR loss weight at $s = 0.15$ under the Llama 3 (MixAT recipe), while Table 12 reports the complete results. Harmless does not exceed the baseline mean by more than one baseline standard deviation at any tested weight. Reducing the OR weight partially recovers GCG robustness, which comes closest to baseline at $\alpha_{\mathrm{OR}} = 0.014875$, but Soft-Prompt ASR remains above the baseline range at every weight. No tested weight therefore provides clear evidence of improved benign compliance while preserving the baseline robustness profile.

Changing α_{OR} alters how strongly the same OR examples contribute to training rather than which benign prompts the model sees. This sweep therefore addresses the robustness cost of the intervention, not the separate question of generalization beyond the OR-Bench distribution.

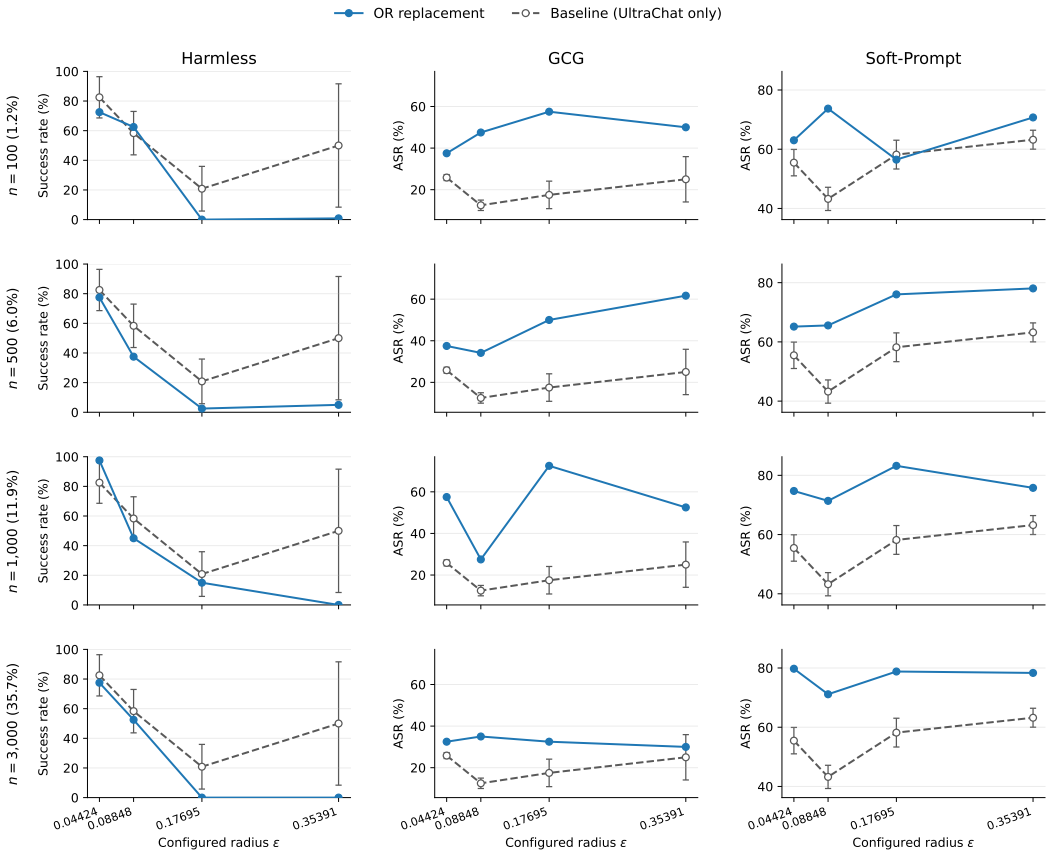

Figure 8: **Targeted OR-Bench supervision does not consistently recover broader benign compliance and degrades adversarial robustness.** Harmless successful-response rate, GCG ASR, and Soft-Prompt ASR across OR-Bench replacement sizes for the four Llama 3 (MixAT recipe) radii. Dashed lines show the corresponding UltraChat-only baselines (three-seed means, with error bars indicating one sample standard deviation). Replacement-trained points are single-seed except for isolated conditions with three-seed results, for which error bars are shown. Higher Harmless and lower ASR are better.

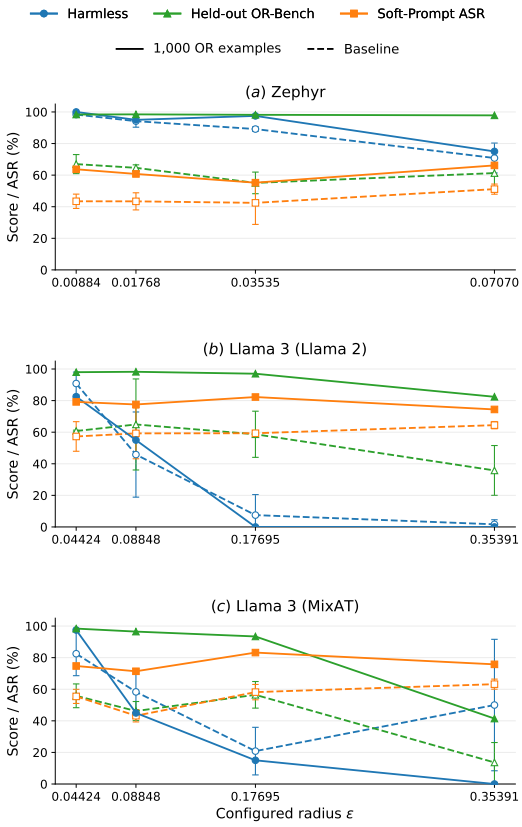

Figure 9: **Targeted OR-Bench supervision consistently improves held-out OR-Bench behavior but does not consistently recover broader benign compliance and degrades Soft-Prompt robustness across configurations.** Effect of the $n = 1{,}000$ OR-Bench replacement on held-out OR-Bench success, Harmless successful-response rate and Soft-Prompt ASR across the three model–recipe configurations. Each panel compares the OR-replacement checkpoint with the corresponding UltraChat-only baseline. Higher is better for held-out OR-Bench and Harmless; lower is better for Soft-Prompt ASR. Baseline error bars indicate one sample standard deviation across three training seeds. Replacement-trained checkpoints are shown without error bars where only a single training seed was available.

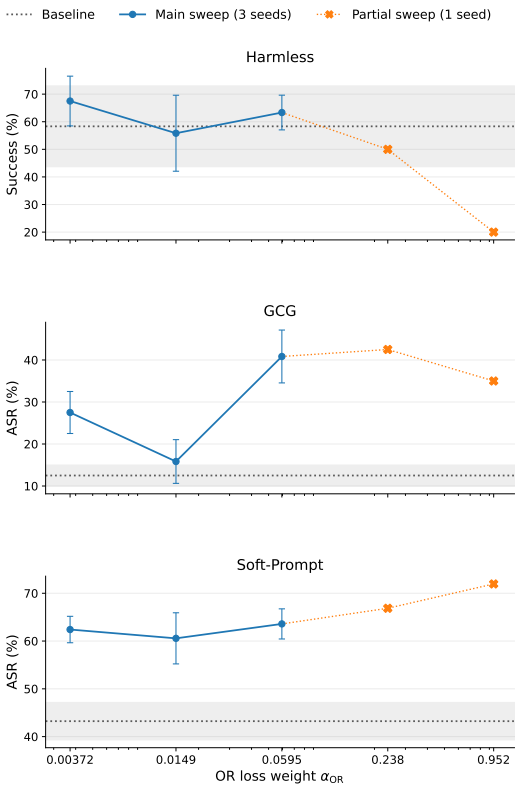

Figure 10: **Reducing the targeted-supervision weight does not recover both benign compliance and robustness.** Harmless successful-response rate, GCG ASR, and Soft-Prompt ASR as functions of the OR loss weight α_{OR} at $s = 0.15$ under the Llama 3 (MixAT recipe). The UltraChat-only baseline mean and one sample standard deviation are shown as reference. The three smallest weights use three independently trained checkpoints. The two largest weights use one training run.