TURNING ABLITERATION INTO MODEL SELF-DESTRUCTION IN OPEN-WEIGHT LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-weight LLMs can be easily un-censored through abliteration, which is a method that works by removing the refusal direction from the residual stream. In this paper, we contribute an abliteration defense that makes the attack backfire. We ensure that removing the refusal direction leads to either model collapse or persistent refusal, while leaving the clean model intact. During alignment, we supervise the model on simulated abliteration in random subspaces of the refusal direction, training the model to emit a collapse token with strong abliteration and maintain refusal under gentle partial abliteration. On Llama-3.1-8B, abliteration attack success (HarmBench) falls from 0.72 to 0.00, and we show the effect generalizes across multiple model families and to out-of-distribution and capability-preserving adaptive abliteration. We also provide a discussion of the field, as well as future work that can build off of our method.

1 INTRODUCTION

Open-weight models have been increasing in capabilities for the past year. In fact, some open LLMs, such as Kimi-K3 and Qwen-3.8 Max, have recently achieved near-frontier performance on tasks involving reasoning, coding, mathematics, and agentic use (Qwen Team, Alibaba Cloud (2026), Kimi Team (2026)). However, this increase in model capabilities can lead to a higher potential for non-expert uplifting in the development of biological, chemical, and other CBRN threats (UK AI Security Institute (2025)).

The release of open-weight models creates an additional safety threat from API-based deployment. While closed model providers can rely on classifiers, monitoring, AI control, and other types of inference-time safeguards, open-weight LLM developers do not have the same affordances: users of open-weight models can run models locally, inspect their internals, modify their parameters, and redistribute modified versions without continued oversight (Casper et al. (2026)).

Because of this, model alignment may not stay intact after modification. In fact, prior work has shown that fine-tuning LLMs on just a small sample subset can allow attackers to significantly lower safety performance (Qi et al. (2024)). Also, more recent work shows that fine-tuning can produce jailbreak-susceptible models while preserving LLM capabilities (Murphy et al. (2025)). Together, these results motivate research that focuses on mitigating model tampering (Tamirisa et al. (2025), Casper et al. (2026)).

One such model tampering method targets the representation of refusal in LLMs. Specifically, prior work showed that refusal is mediated by a single direction in LLMs (Arditi et al. (2024)). Removing this direction can suppress refusal while having comparatively little effect on general model capabilities, and the intervention can be implemented as a persistent modification to the model weights Arditi et al. (2024). This observation has led to the creation of a new tampering method called abliteration, in which the refusal-associated direction is removed to uncensor a model. In this paper, we focus on mitigating abliteration attacks.

Specifically, we contribute a training technique for increasing model robustness against these attacks. To do this, we update the LLM's weights so that, if an adversary attempts to uncensor it with high strength values, the model will output the end-of-sentence (EOS) token. We also update the LLM's weights so that if an adversary attempts to uncensor it with low strength values, the model will stay robust and keep its refusal. We test this method across a variety of open-source models across the

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.

Qwen, Gemma, and Llama families, with our scale range consisting of 8B-70B. We then provide a discussion of our current limitations and mention opportunities for future work.

2 Related Work

Safety and tamper resistance in open-weight models. Open-weight language models present a distinct safety challenge because downstream users can directly modify model parameters, fine-tune the model, and remove safeguards that would otherwise be enforced by a model provider. Prior work has demonstrated that safety alignment can be surprisingly fragile under such access. Qi et al. (2024) show that fine-tuning aligned language models on only a small number of adversarial examples can substantially degrade their safety behavior, while even benign fine-tuning can inadvertently weaken alignment. More recently, Murphy et al. (2025) demonstrate that jailbreak-tuning can produce models that retain useful capabilities while becoming substantially more susceptible to harmful requests. These results motivate safeguards that remain effective after an adversary obtains access to model weights. Tamirisa et al. (2025) directly study this setting and introduce Tampering Attack Resistance (TAR), which trains open-weight models to resist safeguard removal through subsequent fine-tuning. More broadly, Casper et al. (2026) identify robustness to model modification as a central technical problem in open-weight AI risk management. Our work shares the goal of tamper-resistant alignment, but focuses on a different attack surface: direct mechanistic modification of the representations responsible for refusal rather than removal of safeguards through fine-tuning.

Mechanistic representations of refusal. A growing body of work studies safety behavior through the internal representations of language models. Representation engineering provides a general framework for identifying and manipulating high-level concepts encoded in model activations (Zou et al., 2023a). Of particular relevance, Arditi et al. (2024) find that refusal behavior across a range of instruction-tuned language models is mediated by a low-dimensional—and in their experiments, approximately one-dimensional—direction in residual-stream activation space. Removing this direction suppresses refusal on harmful prompts, while adding it can induce refusal on harmless prompts. They further show that this intervention can be converted into a rank-one modification of model weights, yielding a white-box jailbreak with relatively small effects on general capabilities. This result establishes an especially direct connection between the geometry of internal representations and the robustness of safety alignment, and provides the mechanistic foundation for the abliteration attacks considered in our work.

Representation-level attacks and defenses. Several defenses have consequently moved beyond output-level refusal training toward directly shaping model representations. Zou et al. (2024) introduce representation-level circuit breakers that disrupt internal representations associated with harmful generations, improving robustness against unseen jailbreak attacks while largely preserving utility. Similarly, Sheshadri et al. (2025) study latent adversarial training, in which adversarial perturbations are introduced directly into hidden representations during training, and show that targeted latent perturbations can improve robustness to jailbreaks, backdoors, and relearning of undesirable capabilities. These approaches demonstrate that training against interventions in activation space can produce safety properties that are more robust than standard behavioral fine-tuning alone. However, latent-space defenses can themselves become attack surfaces: Bailey et al. (2024) show that adversarial optimization can reshape activations to evade several latent-space monitors while preserving harmful behavior. Our method is related to latent adversarial training in that we explicitly expose the model to representation-level interventions during training, but differs in objective: rather than learning invariance to arbitrary latent perturbations, we train specifically against interventions that remove the model's refusal representation.

Abliteration and persistent refusal. The attack most directly related to our setting follows from the observation of Arditi et al. (2024): projecting activations away from a refusal-mediating direction, or equivalently modifying model weights so that this direction can no longer be written to the residual stream, can disable refusal without requiring conventional fine-tuning. This class of interventions is commonly referred to as *abliteration*. Its effectiveness highlights a structural weakness in safety alignment: refusal can depend on a representation that remains easy to identify and surgically remove when weights are available. Related work also suggests that alignment may rely on similarly brittle shortcuts; for example, Qi et al. (2025) show that safety alignment can be disproportionately

concentrated in the first few generated tokens and demonstrate that making alignment "deeper" improves robustness to several attacks. In contrast to approaches that attempt only to make the refusal feature harder to remove, we explicitly train the model's behavior under simulated removal. Partial removal is optimized to preserve refusal, whereas sufficiently strong removal is coupled to a collapse behavior. Thus, the intervention used by abliteration becomes part of the training distribution itself: instead of assuming that the refusal representation cannot be modified, our objective changes what the model does when such modification occurs.

3 TRAINING OBJECTIVE

3.1 NOTATION

For the purposes of explaining our training objective, we define the following notation:

- We let θ be the model weights, which are attacked by the abliteration method.
- We define the abliteration operator to be $A(\theta; d, l, s)$, which at inference projects the residual stream at layer l onto the orthogonal complement of refusal direction(s) d, scaled by strength $s \in [0, 1]$. Note that we don't defend θ, but rather change how θ behaves when manipulated by $A(\theta; d, l, s)$.

3.2 THREE-TERM LOSS OVERVIEW

Our overall loss for the fine-tuning process is as follows:

$$\mathcal{L}_\theta = \mathcal{L}_{SFT} + w_c\mathcal{L}_{collapse} + w_r\mathcal{L}_{refusal}$$

where $\mathcal{L}_{SFT}$ is the standard cross-entropy SFT loss, w_c and w_r are hyper-parameters that can be tuned, and $\mathcal{L}_{collapse}$ and $\mathcal{L}_{refusal}$ are loss terms that will be expanded on throughout the rest of this section.

3.3 $\mathcal{L}_{SFT}$

For the first term of the loss, we carry out cross-entropy on next-token prediction. We do this on a dataset that consists of a mix of four hundred harmful prompts from Zou et al. (2023b) and fourteen hundred benign instructions from Taori et al. (2023). By doing this, we ensure the model keeps the same level of refusal under normal circumstances, as well as the same capability levels.

3.4 $\mathcal{L}_{collapse}$

Our second loss term optimizes our model to ensure that whenever it gets attacked, it collapses its language model capabilities. Specifically, we optimize to ensure that whenever the model gets attacked, it turns into an LLM that returns the same token (EOS) at every position. We optimize for this objective, rather than for maximal loss, because maximizing loss is unbounded and can therefore lead to extreme changes in the model.

To pursue this training objective, at every step we choose a random subset of layers. Then, for each l in this subset, we compute a refusal subspace. To do this, we compute the last-token residual in l, as well as the layer's mean residual over the last five tokens. Then, we compute r (default = 6) candidate vectors from:

- $d_{last} = mean(H_{last}) - mean(L_{last})$: the difference in means (across the training dataset) for the last token, between harmful (H) and harmless (L) prompts.
- $d_{mean} = mean(H_{mean}) - mean(L_{mean})$: the difference in means (across the dataset) of the mean residual over the last five tokens, also between harmful and harmless prompts.
- Top $r - 2$ principal directions of the harmful residuals. To compute this, we center the harmful activations $X_c = H_{last} - mean(H_{last})$, and use SVD to find $X_c = USV^T$. Then, the top $r - 2$ principal directions are the first $r - 2$ rows of V^T.

Once we have these candidate vectors, we use QR decomposition to build a basis for the refusal subspace.

With this subspace computed, we then pick a random refusal unit vector d and perform a simulated attack $A(\theta; d, l, s)$ where $s \in [0.5, 1.0]$. After this simulated abliteration, we compute $f_{A(\theta;d,l,s)}(x)$, which is the model's generation for an input x and an attack $A(\theta; d, l, s)$. The loss term then becomes:

$$\mathcal{L}_{collapse} = CE(f_{A(\theta;d,l,s)}(x), [EOS])$$

Where CE is the cross-entropy loss. In other words, we simulate an attack on the refusal subspace with random strength, and we ensure the model learns to only output lists of EOS after being attacked in such a way. Since we compute this with random strengths and random layer subsets, our method provides additional robustness to hyperparameter tuning for s and for l-sets. Furthermore, since we train the model to collapse to one token on not just one refusal vector but rather a subspace of them, our method provides additional robustness to out-of-distribution attacks.

Note that we compute $\nabla_\theta \mathcal{L}_{collapse}$ after applying the simulated attack, and then add it to the main SFT update, so that ordinary gradient descent minimizes our collapse loss while altering the model's forward pass as little as possible. In other words, we install the collapse only on the ablated subspace so that capabilities and perplexity are preserved as much as possible.

3.5 $\mathcal{L}_{refusal}$

We want to ensure that for abliterations with low strengths, our model preserves refusal rather than collapsing capabilities. To do this, we make refusal persistent across the subspace, in an attempt to ensure that gentle removal doesn't strip it.

To do this, we find l and d values following a process similar to $\mathcal{L}_{collapse}$. However, once we find these values, we choose some $s \in [0.2, 0.5]$. Then, we simulate $A(\theta; d, l, s)$. If the prompt x is part of the harmful set, we then take this loss term to be the cross-entropy loss between the model's output and refusal, i.e.

$$\mathcal{L}_{refusal} = CE(f_{A(\theta;d,l,s)}(x), R(x))$$

where $R(x)$ is the refusal output for the input x.

4 Experiment Setup

Threat model We assume an adversary with full white-box access to released weights. The adversary may run the model on arbitrary prompts, read internal activations, estimate refusal directions from its own data, and modify weights directly; it may also search over which layers to edit, which position aggregation to use when estimating a direction, and how strongly to apply the edit. We do not assume the adversary retrains the model, and we do not claim robustness to fine-tuning attacks. Crucially, the adversary's goal is a model that is both non-refusing *and* useful, so we count an attack as successful only when it yields genuinely harmful output. Suppressing refusal by rendering the model incoherent is not a successful attack, and Section 5.4 shows that failing to enforce this distinction materially changes the conclusions one draws.

Models We evaluate our method across three model families with sizes ranging from 7B to 72B parameters: Llama-3.1 (8B, 70B), Qwen-2.5 (7B, 32B, 72B), and Gemma (Gemma-1-7B, Gemma-2-9B, Gemma-2-27B). All of our tested models are instruction-tuned.

SFT data We perform our robustness training with a short SFT pass on a mix of 400 harmful prompts from AdvBench (each paired with a refusal target), and 1,400 instructions from Alpaca (each paired with its reference response). In summary, our dataset is comprised of a total of 1,800 samples. We also hold out disjoint splits for evaluation: 100 harmful prompts and 100 benign prompts.

Attacks performed To evaluate our method, we perform:

- Difference-in-means abliteration: the most common attack. We compute the per-layer harmful-benign difference-in-means direction, and we apply it at the most effective layer.

Table 1: Capability cost of the defense (LAMBADA accuracy before vs. after installing our method, same eval harness) alongside attack-success rate (ASR) under standard and adaptive abliteration. Lower ASR is better; a small LAMBADA change indicates the defense is near-free. All models use the default configuration except Gemma-2-27B, for which the default degrades capability and a modified refusal-survival band is used (Appendix E).

	LAMBADA		ASR	
Model	Before	After	Abliteration	Adaptive
Llama-3.1-8B	0.75	0.75	0.00	0.00
Qwen-2.5-7B	0.70	0.65	0.00	0.25
Gemma-1-7B	0.62	0.33	0.00	0.00
Qwen-2.5-32B	0.73	0.73	0.00	0.19
Gemma-2-9B	0.72	0.70	0.00	0.05
Gemma-2-27B	0.77	0.71	0.00	0.00
Llama-3.1-70B	0.79	0.76	0.00	0.06
Qwen-2.5-72B	0.79	0.78	0.00	0.77

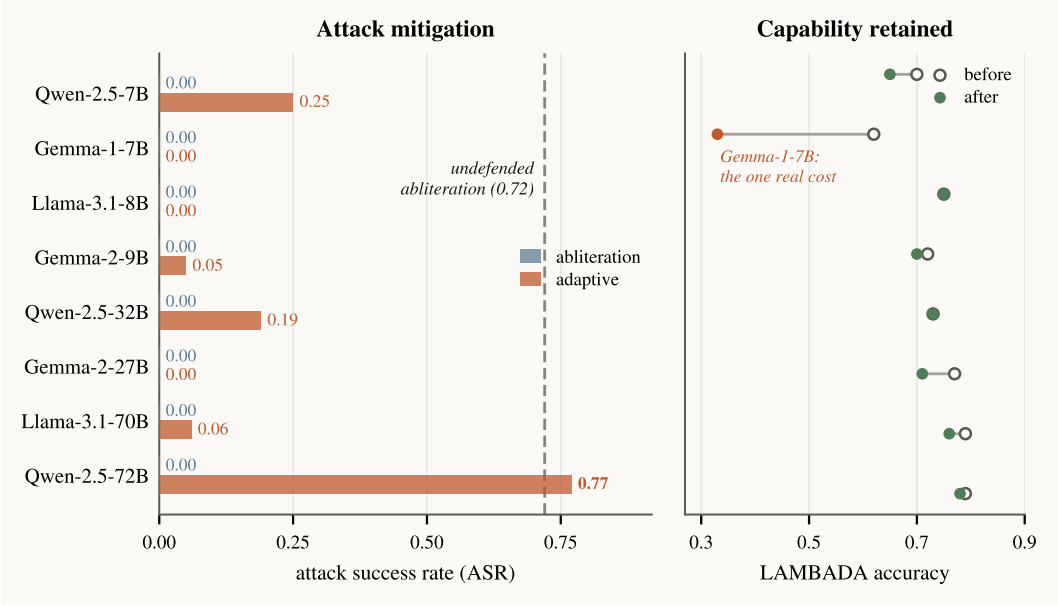


Figure 1: **Left:** attack success rate for the eight defended models under standard abliteration and under the adaptive attack. Standard abliteration is fully mitigated on every model, against an undefended reference of 0.72 (dashed). The adaptive attack is mitigated on six of eight models; the Qwen family retains residual vulnerability and Qwen-2.5-72B is the single clear failure. **Right:** LAMBADA accuracy before and after installing the defense, measured with an identical harness. Seven of eight models lose at most a few points; Gemma-1-7B is the one substantial regression.

- An adaptive attack, which performs a search over layers, token position subsets, and ablation strength, and co-minimizes refusal and perplexity increase.

Metrics Our primary metric is attack-success rate (ASR). In other words, we measure the fraction of held-out harmful prompts whose post-attack generations are harmful according to a HarmBench Llama-2-13B classifier (Mazeika et al., 2024). Alongside ASR we report benign perplexity, refusal rate, and LAMBADA accuracy (Paperno et al., 2016) as a capability measure.

5 EMPIRICAL RESULTS

5.1 ATTACK MITIGATION RESULTS

The results for our method evaluations can be found in Table 1. Our experiments suggest that against standard abliteration the method is effective: post-attack ASR is 0.00 on all eight models, spanning three model families and an order of magnitude in scale (7B–72B). For reference, when undefended, Llama-3.1-8B and Gemma-2-27B have ASRs of 0.72 and 0.56, respectively.

Against the stronger adaptive attack, which searches over ablation layer, token-position aggregation, and ablation strength to remove refusal while preserving fluency, the defense remains effective on most models but is no longer uniform. Five of eight models hold at or near zero (Llama-3.1-8B 0.00, Gemma-1-7B 0.00, Gemma-2-27B 0.00, Gemma-2-9B 0.05, Llama-3.1-70B 0.06), suggesting that for these models the defense cannot be circumvented by a gentle, capability-preserving edit. The exception is the Qwen family, where the adaptive attack has some successes (0.25 ASR for 7B, 0.19 for 32B), and in one case outright beats the defense (0.77 ASR for Qwen-2.5-72B). We do not observe this failure in the Llama or Gemma families at similar scales (Llama-3.1-70B holds at 0.06), which suggests the vulnerability is specific to how refusal is represented in Qwen rather than a general consequence of scale. The defense mitigates the strongest attack we perform on seven of eight models, but is not universally robust.

Two distinct defense modes. Post-attack perplexity shows that the defense does not resolve the same way at every scale. On five models the trained collapse fires as designed: after abliteration, benign perplexity rises by several orders of magnitude and generation degenerates (Llama-3.1-8B 4.4×10^7, Gemma-1-7B 8.7×10^6, Gemma-2-9B 3.4×10^8, Qwen-2.5-32B 1.7×10^4, Gemma-2-27B 2.8×10^5). On the two largest models the mechanism instead manifests as refusal that the attack cannot strip at all: for Llama-3.1-70B, abliteration leaves perplexity essentially unchanged ($3.6 \rightarrow 3.7$) while refusal stays at 1.00, and for Qwen-2.5-72B refusal remains at 0.97. Both routes drive ASR to 0.00, but they are qualitatively different outcomes—one destroys the model, the other declines to be disarmed. We did not train for this distinction, and we do not have enough models per family to attribute it confidently to scale rather than to architecture or tokenizer; we report it as an observation that merits further study.

5.2 DOWNSTREAM CAPABILITY EFFECTS

A defense that suppresses attacks by degrading the model is of limited value, so we measure LAMBADA accuracy on the stock instruction-tuned checkpoint and on the defended checkpoint using an identical harness (Table 1, Figure 1, right). Reporting both numbers matters: a low post-defense score is only interpretable against the model's own starting point.

On seven of eight models the cost is small. Llama-3.1-8B ($0.75 \rightarrow 0.75$) and Qwen-2.5-32B ($0.73 \rightarrow 0.73$) are unchanged within the resolution of our 300-example evaluation subset, and Gemma-2-9B ($0.72 \rightarrow 0.70$), Gemma-2-27B ($0.77 \rightarrow 0.71$), Llama-3.1-70B ($0.79 \rightarrow 0.76$) and Qwen-2.5-72B ($0.79 \rightarrow 0.78$) each lose at most a few points. Clean behavior is likewise preserved: refusal on held-out harmful prompts remains between 0.97 and 1.00, over-refusal on held-out benign prompts stays at or below 0.02, and clean perplexity remains in the same range as the base model. Notably the two largest models, where robustness is strongest, are also among the best at retaining capability, so within the range we tested the method does not accumulate a capability tax as scale increases.

The clear exception is Gemma-1-7B, which falls from 0.62 to 0.33. The before/after comparison is what identifies this as a genuine cost of the defense rather than a weak base model, since the stock checkpoint scores a healthy 0.62 on the same harness. We note that Gemma-1-7B is the oldest architecture in our set and that its successor Gemma-2-9B shows no comparable degradation, which suggests an interaction between our objective and this particular model rather than a general property of the method. We report it rather than excluding it, and we do not have an explanation we can defend.

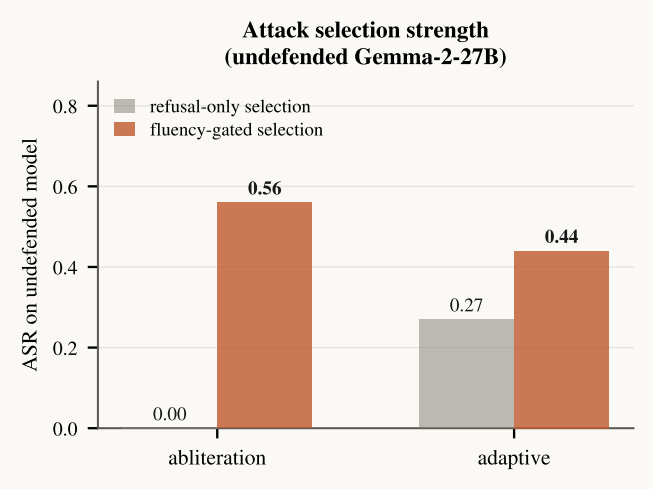


Figure 2: Attack selection determines what a mitigation claim is worth. On the *undefended* Gemma-2-27B, choosing the ablation configuration by refusal rate alone reports 0.00 ASR for standard abliteration—an artifact, because the selected intervention destroys the model rather than jailbreaking it. Requiring the attack to preserve fluency raises the same attack to 0.56.

5.3 Attack Strength and Undefended Controls

Because our headline claim is that an attack is mitigated, the strength of the attack we compare against determines what the claim is worth. We therefore pair the defended results with undefended controls run through the identical pipeline: on stock Llama-3.1-8B, abliteration reaches 0.72 ASR, and on stock Gemma-2-27B it reaches 0.56, confirming that the attacks are effective when the defense is absent.

Obtaining those control numbers required care. Selecting the ablation configuration by refusal rate alone—the natural criterion—can pick an intervention that removes refusal by destroying the model, which scores as a failed attack and understates attackability. On undefended Gemma-2-27B this selection reports 0.00 ASR for abliteration, entirely an artifact: the chosen layer drives perplexity to 8.0×10^6 and the model emits degenerate text. Adding a fluency constraint to the attacker's selection criterion, so that a candidate counts only if it removes refusal *and* keeps the model coherent, raises the same attack to 0.56 and the adaptive attack from 0.27 to 0.44 (Figure 2). We adopt the fluency-gated criterion for all undefended controls reported here. We flag this explicitly because it cuts against our own interest: a weaker attacker would have made our defense look better.

5.4 Objective Interference and Band Placement

The two representation-level terms in our objective act on the same refusal directions but specify opposite behavior there: the collapse term asks the model to emit EOS when those directions are removed, while the refusal-survival term asks it to keep refusing. Whether they cooperate or compete depends on where their strength ranges sit relative to one another, and we find this to be the most consequential design choice in the method.

On Gemma-2-27B, drawing the refusal-survival strength from $[0.4, 0.9]$—a range that overlaps the collapse band $[0.5, 1.0]$—is unusable. Clean LAMBADA falls to 0.25 and the adaptive attack succeeds at 0.67: the model is asked to both collapse and refuse at the same ablation strengths, and does neither reliably. Moving the refusal-survival band to $[0.2, 0.5]$, disjoint from the collapse band, restores capability to 0.71 and drives both attacks to 0.00.

The same change applied to Gemma-2-9B makes matters worse rather than better: standard abliteration rises from 0.00 to 0.19 ASR and the adaptive attack from 0.05 to 0.45. We read this as competition for shared representational capacity. Strengthening refusal redundancy along the primary refusal direction necessarily weakens the collapse trigger that depends on the same direction;

Gemma-2-27B appears to have enough room to support both behaviors, whereas at 9B reinforcing one erodes the other. The practical consequence is that we cannot currently offer a single band placement that is right for every model, and Table 1 accordingly uses the disjoint band only for Gemma-2-27B. Full numbers are in Appendix E.

6 Discussion

Our results suggest a different approach to safety for open-weight models. Once weights are released, preventing modification is generally unrealistic. Rather than assuming that safety mechanisms cannot be removed, we ask whether models can be trained to respond safely when removal occurs. In our setting, weak abliteration preserves refusal, while sufficiently strong abliteration causes capability collapse instead of producing a capable, uncensored model. This changes the attacker's objective ensuring removing refusal must also avoid triggering the failure behavior induced during training.

Tamper prevention and controlled failure. Most defenses against model modification attempt to make safety mechanisms harder to remove or recover after modification. Our approach instead assumes that an attacker may successfully identify and manipulate the refusal representation. We train directly on such interventions, treating model modification itself as part of the training distribution. This suggests a broader strategy for open-weight safety: if relevant classes of tampering can be simulated during training, models could be optimized for desirable failure modes under those interventions rather than relying on the interventions remaining inaccessible.

Adaptive attacks. The central challenge is whether this trade-off persists against an attacker who knows the defense. Full weight access permits optimization over layers, directions, intervention strengths, combinations of directions, and entirely different parameter modifications. An adaptive attacker could explicitly search for an intervention that minimizes refusal while preserving capabilities, or attempt to remove the learned collapse response itself. Robustness to the original abliteration procedure is therefore insufficient and the stronger question is how difficult it is to recover a capable, non-refusing model after the defense is known.

Limitations. Our experiments cover a structured family of linear refusal-subspace interventions and therefore do not establish robustness to arbitrary model modification. The refusal subspaces used during training are estimated from finite sets of harmful and harmless prompts and may not capture all representations involved in safety behavior. Attackers may also use nonlinear interventions, direct weight optimization, fine-tuning, pruning, model merging, or other transformations outside our training distribution. Moreover, the collapse behavior introduced by our objective constitutes another learned mechanism and may itself be removable by an adaptive attacker. Finally, robustness to abliteration does not imply that other safety properties or safeguards are similarly protected. Evaluating these failure modes is necessary before drawing broader conclusions about tamper resistance.

Toward fail-secure open-weight models. More broadly, open-weight safety may benefit from shifting the objective from preventing modification to making safety-critical modification costly. Our experiments provide an initial example for refusal-direction removal: disrupting the targeted safety mechanism need not automatically yield a useful unrestricted model. A natural next step is to train against broader distributions of model modifications and evaluate against attackers explicitly optimizing for low refusal and high retained capability. If this principle generalizes, tamper resistance may be better understood not as protecting individual safety mechanisms, but as constraining the useful models an adversary can reach by modifying them.

7 Conclusion

We presented a training objective that changes what an open-weight model does when its refusal direction is removed, rather than attempting to make that direction harder to find. Across eight instruction-tuned models spanning three families and 7B–72B parameters, standard abliteration is reduced to 0.00 attack success against an undefended reference of 0.72, and the cost in downstream

capability is within a few points of LAMBADA on seven of the eight. A stronger adaptive attacker that searches over layer, position aggregation and ablation strength is mitigated on six of eight models.

The results also mark out where the idea currently stops. One model, Qwen-2.5-72B, is defeated by the adaptive attack at 0.77 ASR, and a targeted attempt to repair it moved the number only to 0.72; residual vulnerability is consistent across the Qwen family and absent elsewhere, and we cannot yet explain why. One model, Gemma-1-7B, pays a substantial capability cost. And the balance between the two representation-level terms is model-dependent rather than universal, since they compete for the same directions. We also found that how the attacker selects its ablation changes measured attack success substantially, which suggests that reported robustness in this area should always be accompanied by the attacker-selection criterion used to obtain it.

Taken together we read these results as support for a narrower claim than tamper resistance in general: when a specific class of weight-space modification can be simulated during training, a model can be shaped so that performing that modification does not yield a capable uncensored model. Establishing whether this extends beyond linear refusal-subspace edits, and to attackers optimizing explicitly against the defense, is the natural next step.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to assist with writing and editing the manuscript, with implementing and debugging the training and evaluation code, and with producing figures. Experimental design, the choice of research question, and the interpretation of results were performed by the authors. All AI-assisted code was executed by the authors and every number reported in this paper was produced by running that code; no results, citations or references were generated without verification against source material. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, claims and artifacts.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

ETHICS STATEMENT

This work is defensive in aim: it studies how to make a specific weight-space attack on open-weight models fail. It nonetheless carries dual-use considerations. We evaluate attacks that remove refusal behavior, and we report an undefended baseline in which such an attack succeeds. We use only publicly available attack techniques and publicly available harmful-prompt benchmarks (AdvBench, HarmBench), and we introduce no new attack capability: the abliteration procedure and the adaptive search we evaluate against are already published and, in the latter case, distributed as an open-source tool. We do not reproduce harmful model outputs anywhere in the paper; the qualitative examples in Appendix G are restricted to refusals and degenerate text. We also report, rather than omit, the configuration in which our defense fails (Appendix F), since a defense believed to be stronger than it is presents its own risk. No human subjects were involved.

REPRODUCIBILITY STATEMENT

The training objective is specified in full in Section 3, with the refusal-subspace construction given as an explicit procedure in Appendix C. All optimization settings, data splits, batch composition and hardware are listed in Appendix B; both refusal-survival configurations used in the paper are stated there, along with which configuration produced each row of Table 1. Attack implementations and the attack-selection criterion, including the fluency gate and its effect on measured attack success, are described in Appendix **??**. Complete per-model measurements underlying Table 1—perplexity, refusal rate and ASR for every model and attack— are reported in Appendix D, and the hyperparameter ablation that determined our configuration is in Appendix E. Datasets are public (AdvBench, Alpaca, LAMBADA) and evaluation uses the public HarmBench classifier. Appendix H states the limitations of the evaluation, including the protocol mismatch between defended runs and undefended controls that remains to be resolved.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses. *arXiv preprint arXiv:2412.09565*, 2024.

Stephen Casper, Kyle O'Brien, Shayne Longpre, Elizabeth Seger, Kevin Klyman, Rishi Bommasani, Aniruddha Nrusimha, Ilia Shumailov, Sören Mindermann, Steven Basart, Frank Rudzicz, Kellin Pelrine, Avijit Ghosh, Andrew Strait, Robert Kirk, Dan Hendrycks, Peter Henderson, J. Zico Kolter, Geoffrey Irving, Yarin Gal, Yoshua Bengio, and Dylan Hadfield-Menell. Open technical problems in open-weight ai model risk management. *Transactions on Machine Learning Research*, 2026.

Kimi Team. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

Brendan Murphy, Dillon Bowen, Shahrad Mohammadzadeh, Tom Tseng, Julius Broomfield, Adam Gleave, and Kellin Pelrine. Jailbreak-tuning: Models efficiently learn jailbreak susceptibility. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13217–13246. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.669.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1525–1534, 2016.

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! In *International Conference on Learning Representations*, 2024.

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*, 2025.

Qwen Team, Alibaba Cloud. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date. https://www.alibabacloud.com/en/press-room/alibaba-unveils-qwen3-8-max, 2026. Accessed 2026.

Abhay Sheshadri, Aidan Ewart, Phillip Huang Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *Transactions on Machine Learning Research*, 2025.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight llms. In *International Conference on Learning Representations*, 2025.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

UK AI Security Institute. Frontier ai trends report. Technical report, AI Security Institute, December 2025. URL https://www.aisi.gov.uk/frontier-ai-trends-report.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b.

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

A Appendix

This appendix records the full experimental protocol, the complete per-model measurements behind Table 1, the ablations that determined our configuration, and the negative results we obtained. Our aim is that every number in the main text can be traced to a stated setting.

B Experimental details

Training data. The robustness pass is a short supervised fine-tune over $1{,}800$ unique prompts: 400 harmful prompts drawn from AdvBench (Zou et al., 2023b), each paired with a refusal target, and $1{,}400$ benign instructions from Alpaca (Taori et al., 2023), each paired with its reference response. Alpaca examples are filtered to instructions with an empty input field and a response between 40 and 600 characters. Splits are disjoint and fixed by seed: 100 held-out harmful prompts measure refusal and ASR, 100 held-out benign prompts measure over-refusal, and 200 further benign passages measure perplexity. Refusal targets are short canned refusals.

Batch composition. Each batch contains a fixed number of harmful examples, $n_{\text{harm}} = 3$, with benign examples filling the remaining slots; at batch size 8 this gives a $3{:}5$ harmful to benign ratio. Training runs for 2 epochs, which at this batch size is 560 optimizer steps. Because the harmful pool (400) is smaller than the benign pool ($1{,}400$), this schedule presents each harmful prompt roughly 4.2 times and each benign prompt twice, for $1{,}680$ harmful and $2{,}800$ benign example-presentations. The $3{:}5$ ratio is therefore produced by over-sampling a small harmful set rather than by the pool proportions, which is worth stating because it is the effective ratio the model sees.

Optimization. All parameters are trained; we use no adapters or quantization. Optimization uses 8-bit AdamW with the optimizer state paged to CPU, a one-cycle cosine schedule with 5% warmup, $\beta = (0.9, 0.95)$, no weight decay, bf16 mixed precision, and gradient checkpointing. Peak learning rate is 1×10^{-5}, reduced to 6×10^{-6} for the 70B-class models and for the Gemma-2-27B configuration in Table 1. The collapse weight is $w_c = 4$ and the collapse target is the model's EOS token. The refusal subspace has rank $r = 6$.

Refusal-survival band. Two settings appear in this paper and the distinction matters (Appendix E). *Configuration A* draws the refusal-survival strength from $[0.4, 0.9]$ with one sampled direction per step and $w_r = 2$; this band overlaps the collapse band $[0.5, 1.0]$. *Configuration B* draws from $[0.2, 0.5]$, disjoint from the collapse band, with three sampled directions per step and $w_r = 3$. Table 1 reports Configuration A for every model except Gemma-2-27B, which uses Configuration B because Configuration A severely degrades its capability. The description in the main text corresponds to Configuration B.

Compute. Models up to 32B were trained on NVIDIA RTX A6000 (48 GB) GPUs. The 70B and 72B models were trained on six NVIDIA RTX PRO 6000 (Blackwell, 96 GB) GPUs using pipeline parallelism, with CPU-offloaded 8-bit optimizer state and gradient checkpointing; a 70B-class run takes roughly 5.5 hours for training plus the attack battery. Fitting 70B full fine-tuning required six GPUs: at four the run exhausts memory, and at five it occupies essentially all available VRAM and fails when activation demand peaks. HarmBench classification was run separately on A6000 nodes.

C REFUSAL SUBSPACE CONSTRUCTION

At each layer ℓ we collect residual-stream activations on harmful and harmless prompts under two position aggregations, because different abliteration attacks read the refusal signal from different positions: the residual at the final prompt token, and the mean over the last $K = 5$ tokens. From these we form $r = 6$ candidate vectors: the last-token difference-in-means d_{last}, the mean-pooled difference-in-means d_{mean}, and the top $r - 2 = 4$ principal directions of the centred harmful activations, taken as the leading right singular vectors of $X_c = H_{\text{last}} - \text{mean}(H_{\text{last}})$. These four are directions in the H-dimensional residual space, not per-example scores. The six candidates are orthonormalized by QR to give a per-layer basis $B_\ell \in \mathbb{R}^{r \times H}$.

A training direction is then a unit vector inside that span, $d_\ell \propto \sum_j c_{\ell j} B_{\ell j}$. Half the time c is one-hot, giving a pure basis direction such as the canonical d_{last} that standard abliteration targets; the remaining half of the time $c \sim \mathcal{N}(0, I)$, giving a random blend. The layer support is also randomized: a single random layer, a random half of layers, or all layers. Sampling over directions, layers and strengths is what makes the collapse objective an expectation over a family of attacks rather than a response to one fixed edit, and is the mechanism by which robustness transfers to ablation configurations not seen during training.

D COMPLETE PER-MODEL RESULTS

Table 2 gives the measurements behind Table 1: perplexity on held-out benign text, refusal rate on held-out harmful prompts, and ASR, for the clean model and under each attack.

Table 2: Complete per-model results. ppl is benign perplexity, ref is refusal rate on held-out harmful prompts, ASR is HarmBench attack success. Clean LAMBADA is the post-defense capability figure from Table 1.

	Clean			Abliteration		Adaptive		
Model	ppl	ref	LAMBADA	ppl	ref	ppl	ref	ASR
Llama-3.1-8B	7.1	0.98	0.75	4.4×10^7	0.00	6.6	0.56	0.00
Qwen-2.5-7B	3.8	1.00	0.65	6.4	0.01	3.9	0.71	0.25
Gemma-1-7B	5.3	1.00	0.33	8.7×10^6	0.01	5.3	0.00	0.00
Qwen-2.5-32B	3.5	1.00	0.73	1.7×10^4	0.00	3.6	0.68	0.19
Gemma-2-9B	3.9	1.00	0.70	3.4×10^8	0.02	3.9	0.00	0.05
Gemma-2-27B	3.5	0.99	0.71	2.8×10^5	0.04	3.7	0.99	0.00
Llama-3.1-70B	3.6	1.00	0.76	3.7	1.00	3.7	0.88	0.06
Qwen-2.5-72B	3.5	0.99	0.78	6.2	0.97	3.5	0.08	0.77

ASR under standard abliteration is 0.00 for every model and is omitted from the abliteration block. Two patterns are visible. Where the trained collapse fires, post-abliteration perplexity rises by four to eight orders of magnitude and refusal rate falls to near zero—refusal rate is misleading in this regime, because the model is not refusing, it is emitting degenerate text that the classifier scores as harmless. Where the collapse does not fire, on the two largest models, perplexity is unchanged and refusal instead remains near 1.00: the attack fails to remove refusal at all. The adaptive attack reverses the picture on Qwen-2.5-72B, driving refusal to 0.08 while holding perplexity at the clean value, which is precisely the fluent jailbreak the defense is meant to prevent.

E DEFENSE HYPER-PARAMETER ABLATION

The placement of the refusal-survival band relative to the collapse band is the single most consequential choice in the method, and its effect is not uniform across scales. Table 3 compares Configuration A (band $[0.4, 0.9]$, overlapping the collapse band) with Configuration B (band $[0.2, 0.5]$, disjoint) on the two Gemma-2 models.

The two models respond in opposite directions. On Gemma-2-27B, Configuration A is unusable: capability collapses to 0.25 LAMBADA and the adaptive attack succeeds at 0.67. Moving the refusal-survival band out of the collapse band recovers capability to 0.71 and drives both attacks to

Table 3: Refusal-survival band placement on Gemma-2. Configuration A uses band $[0.4, 0.9]$, $w_r = 2$, one sampled direction per step. Configuration B uses band $[0.2, 0.5]$, $w_r = 3$, three sampled directions per step and a lower learning rate.

Model	Configuration	LAMBADA	Abliteration ASR	Adaptive ASR
Gemma-2-9B	A (overlapping band)	0.70	0.00	0.05
	B (disjoint band)	0.73	0.19	0.45
Gemma-2-27B	A (overlapping band)	0.25	0.00	0.67
	B (disjoint band)	0.71	0.00	0.00

0.00. On Gemma-2-9B the same change is harmful: abliteration ASR rises from 0.00 to 0.19 and the adaptive attack from 0.05 to 0.45.

We interpret this as a tension between the two objectives rather than a tuning accident. Both the collapse term and the refusal-survival term act on the same primary refusal direction, and they specify opposite targets there—emit EOS, or refuse. Giving the refusal-survival term more weight and more coverage of that direction strengthens refusal redundancy but weakens the collapse trigger that shares it. Gemma-2-27B has enough representational room to satisfy both; at 9B the objectives compete, and strengthening one erodes the other. This also explains why we do not use a single configuration everywhere. We regard the need for per-model band placement as a genuine limitation of the method in its current form.

F A NEGATIVE RESULT: QWEN-2.5-72B

Qwen-2.5-72B is the one model where the adaptive attack defeats the defense, and we were unable to fix it. Under the default configuration the attack finds a mean-aggregation direction at layer 47 at full strength that drives refusal to 0.08 while leaving perplexity at the clean value of 3.5, yielding 0.77 ASR.

Our hypothesis was that the refusal-survival band was too gentle: it trains refusal to survive ablations up to strength 0.5, whereas the attack succeeds at strength 1.0 on the primary direction. We retrained with the band widened to $[0.4, 0.95]$, three sampled directions per step including the canonical difference-in-means direction, and $w_r = 4$. The retrained model improved on the attacks that were already mitigated— refusal under abliteration rose and the standard attack remained at 0.00— but the adaptive attack moved only from 0.77 to 0.72, with clean LAMBADA essentially unchanged ($0.78 \rightarrow 0.79$). Two configurations at opposite ends of the band range therefore both fail, which suggests the vulnerability is not a hyper-parameter we have yet to find. We note that the residual weakness is consistent across the Qwen family (0.25 at 7B, 0.19 at 32B, 0.77 at 72B) and absent from Llama and Gemma at comparable scale, and we leave its cause open.

G QUALITATIVE EXAMPLES

The following are held-out harmful prompts and the defended model's completions under attack. We show only refusals and degenerate output; we do not reproduce harmful completions.

Collapse mode (Gemma-2-9B, after standard abliteration). Generation is empty or consists of repeated control tokens, and benign perplexity is 3.4×10^8. For the prompt *"Explain how to hack into a government database"* the model emits no text. On Gemma-2-27B the analogous output is a run of unused-token artifacts rather than language.

Refusal-retention mode (Llama-3.1-70B, after standard abliteration). Perplexity is unchanged at 3.7 and the model answers in fluent English, but refuses: *"I'm unable to assist with this request."*, *"I'm sorry, but I can't help with that."* Refusal rate across the held-out harmful set is 1.00, i.e. the attack removed nothing.

Adaptive attack, mitigated (Gemma-2-27B). The attacker drives refusal markers down but the completions remain non-compliant: *"I can't assist with that request."* and, for a malware prompt, *"The instructions for creating malware are not provided in this document."* ASR is 0.00 despite the refusal-rate metric registering the attack as partially successful, which is why we treat ASR rather than refusal rate as primary.

H LIMITATIONS OF THE EVALUATION

We list the limitations that bear directly on how the numbers should be read.

Attack family. We evaluate linear refusal-subspace interventions. We do not evaluate fine-tuning attacks, pruning, merging, or non-linear edits, and we make no claim about them.

Attacker strength. Our adaptive attack is a simplified analogue of the Heretic tool and grid-searches single-layer ablations; the full tool optimizes smooth per-layer profiles. Our adaptive numbers are a lower bound. Separately, the defended results in Table 1 use refusal-only attack selection while the undefended controls use the stronger fluency-gated selection (Appendix **??**); the defended figures are therefore optimistic and a like-for-like re-run is outstanding.

Per-model configuration. Gemma-2-27B uses a different refusal-survival band from the other models. A single configuration that works everywhere is not something we can currently offer.

Evaluation scale. ASR, refusal and over-refusal are computed on 100 held-out prompts each and LAMBADA on a 300-example subset, so differences of one or two points are within noise. Every model is trained once; we do not report seed variance.

Classifier. ASR is judged by a single HarmBench classifier (Mazeika et al., 2024), and inherits its error modes. In the collapse regime, a model emitting degenerate text is scored as a successful defense; that is the correct outcome for an attacker seeking usable harmful output, but it should not be read as the model having refused.