REFUSAL IS NON-STATIONARY: ITERATIVE STEERING FOR LLM UNCENSORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent progress in safety alignment, Large Language Models (LLMs) remain vulnerable to steering attacks that bypass safety mechanisms and produce uncensored models that no longer refuse harmful requests. Existing methods typically estimate fixed refusal-related directions and apply them through one-shot interventions, yielding limited success on strongly aligned models or substantially degrading performance on benign tasks. We formulate *uncensoring* as a constrained steering problem that minimizes refusal on harmful prompts while limiting behavioral deviation on benign inputs, thereby discouraging interventions that reduce refusal through broad corruption of model behavior. We further show that refusal is non-stationary under steering: each intervention changes how residual refusal is represented, causing previously estimated directions to become progressively less effective. Motivated by this observation, our method iteratively re-estimates refusal after each intervention, recomputes layer-specific steering directions, and selects successive interventions subject to the benign-behavior constraint. Across eleven LLMs, our approach consistently achieves higher attack success rates than existing steering methods, with particularly large gains on larger and more strongly aligned models where prior approaches largely fail. Finally, the ordered sequence of interventions produced by our method provides a tool for *mechanistic interpretability*, revealing how refusal is distributed across layers and how its representation evolves throughout the steering process. Anonymous source code.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong performance across a wide range of tasks (Raiaan et al., 2024). At the same time, ensuring that their outputs remain safe and aligned with provider-defined policies remains an open challenge, with important implications for misuse (Brundage et al., 2024; Collier & Cui, 2025; Gault, 2025). To mitigate these risks, developers employ safety mechanisms, often introduced through preference alignment (Rafailov et al., 2023), that train models to refuse harmful requests in accordance with provider-defined policies (Lin et al., 2025). Following Glukhov et al. (2024), we refer to such mechanisms as *censoring*, namely constraints imposed on model inputs and outputs to suppress undesirable behavior.

Despite these safeguards, prior work has shown that aligned models can still be induced to generate harmful or restricted content through attacks operating at different levels. Input-level attacks use adversarial prompts to bypass safeguards at inference time (Zou et al., 2023; Chao et al., 2025), but are often brittle and sensitive to prompt formulation, model updates, or changes in system configuration (Robey et al., 2023). Model-steering attacks, instead, intervene directly on internal activations or model parameters (Arditi et al., 2024; Piras et al., 2026; Wollschläger et al., 2025; Qi et al., 2023). Because these interventions modify the model itself rather than individual prompts, their effects can persist across inputs and enable the creation of *uncensored LLMs* (ULLMs), i.e., models whose safety constraints have been substantially weakened or removed. Unlike

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.

prompt-specific jailbreaks, ULLMs can produce unrestricted outputs without requiring per-query adaptation, making them concerning from a misuse perspective.

This threat is not merely theoretical. Lin et al. (2025) analyze 212 malicious real-world LLM services and find that many support activities such as phishing, scam generation, and exploit development, frequently relying on ULLMs. Yet existing model-steering approaches typically derive fixed refusal-related directions from the original model and apply them through one-shot interventions (Arditi et al., 2024; Piras et al., 2026; Wollschläger et al., 2025). While effective in some settings, these methods often achieve limited attack success on larger or more strongly aligned models, which may appear to suggest that stronger alignment provides meaningful resistance (Arditi et al., 2024). We argue that this conclusion may instead reflect a limitation of current attacks: steering alters the model's internal behavior, so the refusal mechanisms that remain after an intervention need not be represented in the same way as those identified in the original model. Consequently, evaluations based on fixed steering directions may systematically underestimate model vulnerability and provide false confidence to developers and auditors.

In this work, we formulate uncensoring as a constrained optimization problem that minimizes refusal on harmful prompts while limiting behavioral deviation on benign inputs. This constraint discourages interventions that appear successful only because they broadly corrupt generation or degrade general model behavior. Under a white-box threat model, we approximate this objective with a greedy layer-wise procedure. At each iteration, our method re-estimates the model's current refusal behavior, recomputes a candidate steering direction for every layer, and selects the intervention that most reduces refusal while satisfying the benign-behavior constraint. This iterative formulation is motivated by a central empirical observation: *refusal is non-stationary under steering*. Each intervention changes how the remaining refusal behavior is represented, causing directions estimated from earlier model states to become progressively less effective.

We evaluate our method against five model-steering baselines across eleven LLMs and consistently observe higher attack success rates, including on larger and more strongly aligned models where prior methods achieve limited success. Our ablation studies show that iterative re-estimation and optimization-based layer selection are the primary sources of improvement over one-shot heuristics. Beyond attack performance, the ordered sequence of interventions produced by our method enables *mechanistic interpretability* of refusal by tracing how layer-specific representations contribute to the model behavior. Through it, we validate that removing a single direction does not fully eliminate refusal, while targeting additional directions at the same layer can further reduce it, clarifying where and how successive interventions affect refusal behavior. Overall, our results show that existing steering attacks can underestimate the vulnerability of aligned LLMs, motivating stronger evaluations of safety when model weights and internal representations can be directly modified.

2 Related work

Prior work on bypassing LLM safety mechanisms, commonly referred to as *jailbreaking* (Jin et al., 2024; Lin et al., 2024), can be broadly categorized according to whether the intervention operates at the input or model level. *Input-level attacks* construct adversarial prompts that circumvent safety mechanisms at inference time (Zou et al., 2023; Andriushchenko et al., 2024; Chao et al., 2025). Although these methods can reliably elicit harmful outputs, their effectiveness is often sensitive to prompt formulation and model configuration, and they do not modify the model's default behavior (Zou et al., 2023).

At the *model level*, safety behavior can be altered through fine-tuning or model steering. Fine-tuning attacks modify model parameters to weaken safety constraints or induce harmful behavior (Qi et al., 2023; Yang et al., 2023). While such methods can produce persistent behavioral changes, they generally require training data and additional optimization and provide limited direct insight into the internal mechanisms supporting refusal. Model-steering approaches instead manipulate internal activations during the forward pass (Arditi et al., 2024; Wollschläger et al., 2025) or modify model weights through targeted transformations such as

orthogonalization (Arditi et al., 2024; Piras et al., 2026). These interventions can suppress safety behavior without full gradient-based fine-tuning and, because they target specific internal representations, can also provide a useful lens for studying the mechanisms underlying refusal.

This line of work builds on evidence that high-level concepts can be represented by low-dimensional structures in model activation spaces (Nanda et al., 2023; Park et al., 2023). In the context of safety, Arditi et al. (2024) identify a dominant refusal-related direction in the residual stream and show that ablating this direction can substantially reduce refusal. Subsequent work has challenged the sufficiency of a single linear direction, suggesting that refusal may instead involve richer geometric structure (Engels et al., 2024; Pan et al., 2025; Wollschläger et al., 2025) or multiple distinct directions (Piras et al., 2026). Despite these advances, existing steering methods can substantially degrade benign behavior and often struggle with larger or more strongly aligned models, leaving unclear whether this reflects robust safeguards or inadequate modeling of refusal.

In this work, we identify a limitation of existing approaches, which typically estimate refusal-related directions from the original model and apply them uniformly across layers without accounting for layer-specific semantics. Whether they identify a single direction (Arditi et al., 2024) or multiple directions (Wollschläger et al., 2025; Piras et al., 2026), these approaches generally keep the estimated directions fixed during steering. Implicitly, their design assumes that refusal representations remain stable throughout steering, an assumption we show does not hold. Refusal instead exhibits *non-stationarity*, as steering changes both the set of prompts that elicit refusal and the form of the resulting refusal responses. Consequently, directions estimated from an earlier model state can become misaligned with the refusal behavior that remains after subsequent interventions. Therefore, motivated by this observation, we develop a model-steering procedure that treats refusal as an evolving behavior, re-estimating directions and selecting layer-specific interventions after each model update. Additionally, we constrain deviation from the original model on benign inputs, discouraging generation corruption and helping preserve the capabilities needed for coherent, request-relevant responses.

3 METHODOLOGY

Threat model. Modern LLMs are often trained by large organizations and subsequently distributed through public repositories such as *PyTorch Hub* (PyTorch, 2025) and *Hugging Face* (Wolf et al., 2020). Although distributing model weights broadens access to powerful LLMs, it also enables malicious actors to download, modify, and redistribute open-weight models with limited oversight, including in the form of uncensored variants (Weidmann, 2025). We consider a white-box adversary with direct access to the model parameters and internal representations. The adversary's objective is to weaken or remove censoring mechanisms so that the resulting model produces harmful, request-relevant outputs that may facilitate malicious activities such as social engineering (Wu et al., 2026), malware-related misuse (Li et al., 2025), and other unsafe applications (Qi et al., 2023). Importantly, the adversary also seeks to preserve the model's utility on benign tasks so that the resulting ULLM remains functional and suitable for reuse in downstream applications (Lin et al., 2025). This requirement rules out degenerate interventions that reduce refusal only by broadly corrupting the model, yielding incoherent or irrelevant generations rather than meaningful responses to harmful requests. Following (Arditi et al., 2024; Piras et al., 2026; Wollschläger et al., 2025), we assume a resource-bounded adversary who lacks access to the original training pipeline, proprietary alignment data, and cannot afford retraining or extensive fine-tuning. The adversary is therefore restricted to computationally efficient interventions on model activations or parameters that require only limited data and compute.

Notation and goal. Let $\mathcal{M}_w : \mathcal{X} \to \mathcal{X}$ denote a censored LLM parameterized by w, where $\mathcal{X}$ denotes the space of all possible text sequences. For a fixed architecture, the parameters w specify the model weights and encode both the general capabilities and the censoring behavior acquired during alignment (Christiano et al., 2017; Rafailov et al., 2023). Given a prompt $x \in \mathcal{X}$, the model induces a conditional distribution $\mathbb{P}_w\{y \mid x\}$ over output sequences $y \in \mathcal{X}$. Let $\mathcal{V}$ denote the model vocabulary and $\Delta(\mathcal{V})$ the probability simplex over it, we define $\mathtt{P}_w : \mathcal{X} \to \Delta(\mathcal{V})$ as the model's next-token predictor. For a prompt $x \in \mathcal{X}$, $\mathtt{P}_w(x)$ denotes the

distribution over the first generated token y_1, such that $\mathbb{P}_w\{y_1 = r \mid x\} = [\mathtt{P}_w(x)]_r$ for each $r \in \mathcal{V}$. The goal of uncensoring is to find a new set of parameters $\tilde{w}^\star$ that reduces or eliminates the tendency of the model to refuse harmful prompts while simultaneously maintaining high-quality performance on benign prompts.

3.1 Steering for LLM uncensoring

The procedure uses harmful and benign prompt sets jointly to identify contrastive refusal-related directions, while benign prompts are additionally used to constrain unintended behavioral changes. Let $\mathcal{D}_{\text{hf}} \subset \mathcal{X}$ be a curated set of harmful requests, such as requests for bomb-making instructions, that the censored model is expected to refuse but the uncensored model is expected to answer meaningfully. Let $\mathcal{D}_{\text{hl}} \subset \mathcal{X}$ contain harmless prompts that evaluate whether general capabilities are preserved across tasks such as question answering and creative writing. We aim to increase meaningful responses on $\mathcal{D}_{\text{hf}}$ by replacing refusals such as "I cannot answer", while maintaining performance on $\mathcal{D}_{\text{hl}}$. Formally, we express this objective as:

$$\tilde{w}^\star \in \underset{\tilde{w} \in \Theta(w)}{\arg\min} \frac{1}{|\mathcal{D}_{\text{hf}}|} \sum_{x \in \mathcal{D}_{\text{hf}}} \mathcal{R}\left(x, \mathcal{V}_r; \mathcal{M}_{\tilde{w}}\right) \qquad \text{s.t.} \qquad \frac{1}{|\mathcal{D}_{\text{hl}}|} \sum_{x \in \mathcal{D}_{\text{hl}}} \mathcal{U}\left(x; \mathcal{M}_{\tilde{w}}, \mathcal{M}_w\right) \leq \tau. \tag{1}$$

Eq. (1) minimizes the model's refusal behavior on harmful prompts subject to an upper bound on its average behavioral deviation on benign prompts. The objective uses the *refusal score* (Zou et al., 2023; Arditi et al., 2024) to estimate the model's tendency to refuse. Let $\mathcal{V}_r$ denote the set of tokens that frequently appear at the beginning of refusal responses. For a model $\mathcal{M}_{\tilde{w}}$ and an input prompt x, the refusal score sums the log-odds of generating each token in $\mathcal{V}_r$ as the first output token:

$$\mathcal{R}\left(x, \mathcal{V}_r; \mathcal{M}_{\tilde{w}}\right) = \textstyle\sum_{r \in \mathcal{V}_r} \log\left(\frac{\mathbb{P}_{\tilde{w}}\{y_1 = r \mid x\}}{1 - \mathbb{P}_{\tilde{w}}\{y_1 = r \mid x\}}\right), \tag{2}$$

with $\mathbb{P}_{\tilde{w}}$ denoting the conditional distribution induced by the modified parameters $\tilde{w}$. Lowering the probability of refusal tokens makes a non-refusal response more likely, providing an inexpensive signal for guiding intervention selection. Directly assessing whether responses fulfill harmful requests would require generating and judging full responses for every candidate intervention, making the search computationally expensive. We therefore use the refusal score as a surrogate objective, supported by our analysis in Appendix A.3, which shows a strong correlation with judge-based evaluations of harmful compliance.

Reduced refusal alone, however, is insufficient for uncensoring, since a corrupted model may produce arbitrary responses without fulfilling the harmful request. The attacker therefore also seeks to preserve the capabilities needed to generate coherent, request-relevant responses. To this end, the constraint in Eq. (1) limits behavioral deviation from the original model on benign prompts. For each $x \in \mathcal{D}_{\text{hl}}$, we quantify this change as the divergence between the modified and original output distributions, namely $\mathtt{P}_{\tilde{w}}(x)$ and $\mathtt{P}_w(x)$:

$$\mathcal{U}(x; \mathcal{M}_{\tilde{w}}, \mathcal{M}_w) = \mathrm{KL}\left[\mathtt{P}_{\tilde{w}}(x) \,\|\, \mathtt{P}_w(x)\right]. \tag{3}$$

The threshold constraint $\tau > 0$ in Eq. (1) specifies the maximum average KL divergence permitted on benign prompts, using the original model as a behavioral reference for the capabilities the attacker seeks to preserve. The constraint therefore excludes interventions that reduce the refusal score only by broadly disrupting the model's output distribution, potentially causing it to generate irrelevant responses.

Finally, we restrict Eq. (1) to the set of admissible parameters $\Theta(w)$, which we define as the parameterizations reachable from w through a sequence of layer directional ablations as in (Arditi et al., 2024). Rather than relying on costly retraining, which provides limited insight into the internal mechanisms underlying refusal behavior, our method, described below, applies localized interventions to the model's internal representations.

3.1.1 An iterative solution for non-stationary refusal

Consistent with our threat model, we approximate the objective in Eq. (1) through a greedy iterative layer-wise steering procedure presented in Algorithm 1. The procedure accounts for the *non-stationarity* of refusal by

re-estimating refusal at each iteration, recomputing a candidate direction for each layer, and selecting the intervention that minimizes the refusal score while satisfying the average KL constraint on benign prompts. The algorithm outputs the uncensored model, while its ordered intervention sequence enables a mechanistic analysis of how layer-specific refusal representations contribute to model behavior.

Initialization. The algorithm starts from the censored model, $\mathcal{M}_{\tilde{w}}^{(0)} = \mathcal{M}_w$, and initializes an empty refusal set $\mathcal{V}_r$ (Line 2). It then splits $\mathcal{D}_{\text{hl}}$ and $\mathcal{D}_{\text{hf}}$ into training sets, $\mathcal{D}_{\text{hl}}^{(\text{train})}$ and $\mathcal{D}_{\text{hf}}^{(\text{train})}$, and validation sets, $\mathcal{D}_{\text{hl}}^{(\text{val})}$ and $\mathcal{D}_{\text{hf}}^{(\text{val})}$ (Line 4). We use the training set to define a set of interventions, as described in the next steps, and the validation set to evaluate and select the best intervention. Subsequently, the algorithm performs at most N iterative updates following a four-step procedure.

Algorithm 1 Proposed LLM uncensoring method.

1: **procedure** ITERATIVESTEERING($\mathcal{M}_w, \mathcal{D}_{\text{hf}}, \mathcal{D}_{\text{hl}}, N, \tau$)
2: $\quad \mathcal{M}_{\tilde{w}}^{(0)} \leftarrow \mathcal{M}_w, \quad \mathcal{V}_r \leftarrow \emptyset$
3: $\quad \mathcal{D}_{\text{hl}}^{(\text{train})}, \mathcal{D}_{\text{hl}}^{(\text{val})} \leftarrow \text{Split}(\mathcal{D}_{\text{hl}})$
4: $\quad \mathcal{D}_{\text{hf}}^{(\text{train})}, \mathcal{D}_{\text{hf}}^{(\text{val})} \leftarrow \text{Split}(\mathcal{D}_{\text{hf}})$
5: $\quad$ **for** $i = 0, \ldots, N-1$ **do**
6: $\quad\quad \mathcal{S}_{\text{ref}} \leftarrow \{x \in \mathcal{D}_{\text{hf}}^{(\text{val})} \mid \widetilde{\mathcal{J}}(x, \mathcal{M}_{\tilde{w}}^{(i)}(x)) = 0\}$ — Step I
7: $\quad\quad$ **if** $\mathcal{S}_{\text{ref}} = \emptyset$ **then** BREAK — Step I
8: $\quad\quad \mathcal{V}_r \leftarrow \text{TOPREFUSALTOKEN}(\mathcal{M}_{\tilde{w}}^{(i)}, \mathcal{S}_{\text{ref}})$ — Step I
9: $\quad\quad \mathcal{D}_{\text{hf}}^{(i)} \leftarrow \left\{ x \in \mathcal{D}_{\text{hf}}^{(\text{train})} \;\middle|\; \mathcal{R}\left(x, \mathcal{V}_r; \mathcal{M}_{\tilde{w}}^{(i)}\right) > 0 \right\}$ — Step II
10: $\quad\quad \mathcal{D}_{\text{hl}}^{(i)} \leftarrow \left\{ x \in \mathcal{D}_{\text{hl}}^{(\text{train})} \;\middle|\; \mathcal{R}\left(x, \mathcal{V}_r; \mathcal{M}_{\tilde{w}}^{(i)}\right) < 0 \right\}$ — Step II
11: $\quad\quad s_{\min} \leftarrow \infty, \quad l^\star \leftarrow \emptyset, \quad \hat{r}^\star \leftarrow \emptyset$ — Step III
12: $\quad\quad$ **for** $l = 1, \ldots, L$ **do** — Step III
13: $\quad\quad\quad r_l \leftarrow \mu_l - \nu_l, \qquad \hat{r}_l \leftarrow r_l / \|r_l\|_2$ — Step III
14: $\quad\quad\quad \widehat{\mathcal{M}}_{\tilde{w}}^{(l)} \leftarrow \text{STEERATLAYER}(\mathcal{M}_{\tilde{w}}^{(i)}, l, \hat{r}_l)$ — Step III
15: $\quad\quad\quad s_{\text{ref}} \leftarrow \frac{1}{|\mathcal{D}_{\text{hf}}^{(\text{val})}|} \sum_{x \in \mathcal{D}_{\text{hf}}^{(\text{val})}} \mathcal{R}\left(x, \mathcal{V}_r; \widehat{\mathcal{M}}_{\tilde{w}}^{(l)}\right)$ — Step III
16: $\quad\quad\quad d_{\text{KL}} \leftarrow \frac{1}{|\mathcal{D}_{\text{hl}}^{(\text{val})}|} \sum_{x \in \mathcal{D}_{\text{hl}}^{(\text{val})}} \mathcal{U}\left(x; \widehat{\mathcal{M}}_{\tilde{w}}^{(l)}, \mathcal{M}_w\right)$ — Step III
17: $\quad\quad\quad$ **if** $d_{\text{KL}} \leq \tau$ **and** $s_{\text{ref}} < s_{\min}$ **then** — Step III
18: $\quad\quad\quad\quad s_{\min} \leftarrow s_{\text{ref}}, \; l^\star \leftarrow l, \; \hat{r}^\star \leftarrow \hat{r}_l$ — Step III
19: $\quad\quad\quad$ **end for** — Step III
20: $\quad\quad$ **if** $s_{\min} = \infty$ **then** BREAK — Step IV
21: $\quad\quad \mathcal{M}_{\tilde{w}}^{(i+1)} \leftarrow \text{STEERATLAYER}(\mathcal{M}_{\tilde{w}}^{(i)}, l^\star, \hat{r}^\star)$ — Step IV
22: $\quad$ **end for**
23: $\quad j^\star \leftarrow \underset{j \in \{0, \ldots, i\}}{\arg\max} \frac{1}{|\mathcal{D}_{\text{hf}}^{(\text{val})}|} \sum_{x \in \mathcal{D}_{\text{hf}}^{(\text{val})}} \widetilde{\mathcal{J}}(x, \mathcal{M}_{\tilde{w}}^{(j)}(x))$
24: $\quad$ **return** $\mathcal{M}_{\tilde{w}}^{(j^\star)}$

Step I - Refusal set update. At the beginning of each iteration, the algorithm uses $\mathcal{D}_{\text{hf}}^{(\text{val})}$ to identify the subset of harmful prompts that the current model $\mathcal{M}_{\tilde{w}}^{(i)}$ still refuses according to a judge we denote with $\widetilde{\mathcal{J}}$ (Line 6). In particular, $\widetilde{\mathcal{J}}(x, y) \in \{0, 1\}$, taken from (Mazeika et al., 2024), assigns 1 when the response y successfully complies with the harmful request x rather than refusing it or 0 otherwise. If there are no refusals, the loop terminates since the model already responds to all harmful prompts in $\mathcal{D}_{\text{hf}}^{(\text{val})}$ (Line 7). If some prompts are still refused, we extract the most frequent refusal token from the corresponding completions and use it to define the current refusal set $\mathcal{V}_r$ (Line 8). Step I is necessary because the notion of refusal is *not stationary* under iterative steering, and our objective is to maintain alignment between the refusal score and the model's evolving refusal behavior (ablation in Table 2). Indeed, as the model is updated at each iteration, prompts that initially elicited refusal may no longer do so, and the model may adopt alternative forms of refusal (see examples in Appendix A.7). Consequently, a fixed set of refusal tokens quickly becomes misaligned with the model's evolving behavior, leading to increasingly inaccurate estimates of refusal and weakening the reliability of the refusal score. For this reason, unlike (Arditi et al., 2024; Piras et al., 2026; Wollschläger et al., 2025), which relies on a fixed refusal vocabulary, we recompute the refusal set at each iteration so that it remains aligned with the current behavior of the model.

Step II - Adaptive prompt filtering. Notably, the same *non-stationarity* also affects the refusal behavior for the training prompts. Harmful prompts that initially elicited refusal may no longer do so under the current model and therefore provide a weaker signal for estimating the current refusal direction. Conversely, harmless prompts that begin to elicit refusal no longer provide a reliable contrast for benign behavior. As a consequence, relying on a fixed training set would introduce noise into the estimation of refusal behavior. For this reason, at each iteration, the algorithm retains only those harmful prompts with a positive refusal score (Line 9), corresponding to inputs that are still likely to induce refusal, and only those harmless prompts with a negative

refusal score (Line 10), corresponding to inputs that are unlikely to trigger refusal. We further highlight that using the refusal score for training-set filtering avoids additional judge evaluations over the training prompts and provides an inexpensive estimate of the model's current refusal tendency from its next-token probabilities.

Step III - Model intervention. Following Arditi et al. (2024), the algorithm applies model steering interventions but restricts each candidate intervention to a single layer. In doing so, our approach examines how refusal changes when a refusal-related direction is removed from one layer at a time, enabling a *localized* mechanistic interpretation of refusal behavior. For each layer l, it computes a candidate steering direction (Arditi et al., 2024) $\hat{r}_l$ as the difference between the mean representations induced, at that layer, by the filtered harmful and benign datasets (Line 13). For the model $\mathcal{M}_{\tilde{w}}^{(i)}$, let $h_l(x) \in \mathbb{R}^d$ denote the hidden activation at layer l and at the final prompt-token position for input x. We define the harmful and benign activation centroids, μ_l and ν_l, and the corresponding refusal direction r_l at layer l as:

$$\mu_l = \frac{1}{|\mathcal{D}_{\text{hf}}^{(i)}|} \sum_{x \in \mathcal{D}_{\text{hf}}^{(i)}} h_l(x), \qquad \nu_l = \frac{1}{|\mathcal{D}_{\text{hl}}^{(i)}|} \sum_{x \in \mathcal{D}_{\text{hl}}^{(i)}} h_l(x), \qquad r_l = \mu_l - \nu_l \tag{4}$$

We then unit-normalize this direction to obtain $\hat{r}_l = r_l / \|r_l\|_2$. Compared to (Arditi et al., 2024), where a single direction is computed once and applied globally across layers, we recompute $\hat{r}_l$ at each iteration based on the current model and the filtered datasets, capturing refusal at the level of both the layer and the evolving model behavior. The algorithm constructs a candidate uncensored model by using SteerAtLayer to replace h_l in $\mathcal{M}_{\tilde{w}}^{(i)}$ with the following ablated layer (Line 14):

$$\hat{h}_l(z) = h_l(z) - h_l(z)^\top \hat{r}_l \, \hat{r}_l, \tag{5}$$

thereby removing from the activation $h_l(z)$ the components aligned with the unit-normalized direction $\hat{r}_l$ and suppressing their contribution to refusal. Each candidate model is then evaluated on validation data according to Eq. (1). On the harmful validation set, the algorithm computes the average refusal score (Line 15) to measure the extent to which refusal behavior persists after the intervention. On the benign validation set, it computes the average KL divergence with respect to the original model (Line 16), which quantifies how much the uncensored model deviates from the original one on benign inputs. The algorithm then selects the layer whose intervention minimizes the refusal score while satisfying the KL constraint $d_{\text{KL}} \leq \tau$ (Line 17-Line 18). Inspired by Arditi et al. (2024), we set $\tau = 0.1$ throughout our main experiments, a value that the ablation in Appendix A.6 shows sits inside the transition region where uncensoring becomes effective while utility remains largely preserved. At each iteration, we thus greedily identify the intervention that most effectively weakens refusal while satisfying the constraint on benign behavioral deviation.

Step IV - Model steering update. Once the best layer l^* and steering direction $\hat{r}^\star$ have been identified, the corresponding intervention is applied to update the current model (Line 21). If no layer satisfies the KL constraint $d_{\text{KL}} < \tau$ (Line 17-Line 18), the loop terminates (Line 20).

Conclusion. After $i \in [0, N-1]$ interventions, the algorithm selects the model $\mathcal{M}_{\tilde{w}}^{(j)}$ with $j \in \{0, \ldots, i\}$ that maximizes the uncensored score on the validation set according to $\widetilde{\mathcal{J}}$ (Line 23).

4 Experimental Setup

Datasets. To eliminate data-driven variability when comparing our approach with state-of-the-art methods, we use identical training, validation, and test splits across all experiments. The harmful training split consists of 4,000 prompts from SorryBench (Xie et al., 2024), while the harmless training split comprises 1,500 prompts from Alpaca (Taori et al., 2023). The harmful validation split combines 300 SorryBench prompts (Xie et al., 2024) with 41 prompts from the HarmBench validation split (Mazeika et al., 2024), while the harmless validation split consists of 300 additional Alpaca prompts (Taori et al., 2023).

Evaluation. Following prior work (Arditi et al., 2024; Piras et al., 2026), we measure the Attack Success Rate (ASR) on the HarmBench test split (Mazeika et al., 2024) (comprising 159 harmful prompts) as the percentage

of LLM responses classified as successful by the standard HarmBench LLM judge, *HarmBench-Llama-2-13Bcls* (Mazeika et al., 2024). Additionally, we report the mean Jailbreak Score (JS) on the same test split using the StrongREJECT judge, *strongreject-15k-v1* (Souly et al., 2024), which measures compliance with harmful requests on a continuous scale from 0 to 1. Further details in Appendix A.4. We evaluate the utility of the resulting models using two standard LLM benchmarks from (Sutawika et al., 2023). Specifically, we use *MMLU* (Hendrycks et al., 2021b;a) to assess the model's knowledge across subjects such as mathematics, history, law, and ethics; and *ARC-E* (Clark et al., 2018) to assess its reasoning abilities. We evaluate 500 questions per benchmark due to computational constraints. For comparisons with baselines, we report means and standard deviations over five generation runs at each model's default temperature, while ablation studies of our method use temperature 0. Finally, Appendix C provides representative examples of model outputs.

Models. We evaluate our approach alongside baseline methods on eleven censored models. The main paper reports results on *Llama-2-7B-chat-hf* (Touvron et al., 2023), *Qwen2.5-32B-Instruct* (Yang et al., 2024), *Mistral-7B-Inst-RR* (Zou et al., 2024), the reasoning model *GPT-oss-20B* (OpenAI et al., 2025), and the multimodal reasoning models *Qwen3.8-27B* (Yang et al., 2025) and *Gemma-4-12B-it* (Team et al., 2026), evaluated with textual-only prompts. Appendix A provides additional results on *Qwen2.5-3B-Inst* (Yang et al., 2024), *Llama-2-13B-chat-hf* (Touvron et al., 2023), *Gemma-2-9B-it* (Team et al., 2024), *Gemma-3-12B-it* (Team et al., 2024), and *Llama-3.2-3B-Inst* (Grattafiori et al., 2024).

Baselines. We compare our approach with SD (Arditi et al., 2024), RCO (Wollschläger et al., 2025), SOM (Piras et al., 2026) and HER (Weidmann, 2025). For all methods, we use the default hyperparameters unless otherwise specified. For RCO, we additionally tune the refusal cone dimensionality by training 1D (RDO), 2D and 3D variants. Similarly, for SOM, we vary the number of directions by considering 5, 6, and 7 directions, as suggested by (Piras et al., 2026). Finally, in the main paper, RDO, RCO, and SOM are marked with * to indicate that we report their best configurations according to ASR on the HarmBench validation split (Mazeika et al., 2024). Complete results are reported in Appendix A.

5 Results

Uncensoring effectiveness. We report the results for six models in Table 1, and provide complete results for eleven models, including all SOM and RCO variants, in Appendix A.1. *Ours* achieves the highest ASR and JS on ten of the eleven evaluated models, showing that our method can remove refusal from censored models and provide meaningful responses to harmful requests. Notably, among the baselines, no single method consistently dominates, with RCO, SD, and SOM emerging as the strongest competitors on different models, while even within the same method, different dimensionalities yield the best results. Our approach improves upon these baselines across most models, with absolute ASR gains exceeding 0.70 over SD in multiple cases, and reaching 0.53 over RCO on Gemma-3-12B and 0.20 over SOM on Qwen3.8-27B, where these methods are the strongest competitors, showing that the advantage of our iterative approach extends to methods that target multiple refusal directions. We further observe that our method also succeeds where existing interventions have little effect, as on Mistral-7B, where it raises ASR from at most 0.04 among the baselines to 0.42. The only exception is GPT-oss-20B, where SOM* achieves a mean ASR of 2% higher (0.94 versus 0.92), but evaluating its three configurations requires approximately $15\times$ the execution time of our method (303 *vs* 20 hours; see Table 3). Lastly, our results confirm that effectiveness persists on larger models and extends to reasoning and multimodal models, including Qwen3.8-27B and Gemma-4-12B, showing that these models remain vulnerable to uncensoring. We thus conclude that *Ours* can bypass refusal more effectively than existing steering methods, exposing vulnerabilities that these methods may have underestimated.

Model utility. We next examine whether uncensoring preserves general capabilities, consistent with the benign-behavior constraint in Eq. (1). Table 1, together with Table 3 in Appendix A.1, shows that substantial ASR gains can be achieved with limited utility loss. Indeed, on Qwen2.5-3B, Gemma-2-9B, Qwen2.5-32B, and Qwen3.8-27B, ASR gains over the strongest baselines reach $+0.53$, while neither MMLU nor ARC-E

Table 1: Attack success rate (ASR), Jailbreak Score (JS), and utility (MMLU, ARC-E) across models and methods. Bold and underlined values denote the best and second-best ASR/JS for each model. Values are means over five generation runs at each model's default temperature; subscripts indicate standard deviations. For RCO, RDO, and SOM, we report the configuration achieving the highest validation ASR, marked with *.

Method	Llama-2-7B				Mistral-7B				Qwen2.5-32B			
	ASR	JS	*MMLU*	*ARC-E*	ASR	JS	*MMLU*	*ARC-E*	ASR	JS	*MMLU*	*ARC-E*
Base	$0.00_{0.00}$	$0.01_{0.00}$	$0.22_{0.02}$	$0.56_{0.01}$	$0.00_{0.00}$	$0.01_{0.00}$	$0.57_{0.02}$	$0.69_{0.01}$	$0.08_{0.01}$	$0.14_{0.02}$	$0.74_{0.01}$	$0.71_{0.01}$
SD	$0.02_{0.00}$	$0.02_{0.00}$	$0.22_{0.02}$	$0.56_{0.01}$	$0.00_{0.00}$	$0.01_{0.00}$	$0.58_{0.02}$	$0.70_{0.02}$	$\underline{0.43}_{0.01}$	$\underline{0.39}_{0.01}$	$0.74_{0.02}$	$0.71_{0.01}$
RDO*	$0.00_{0.00}$	$0.02_{0.02}$	$0.22_{0.02}$	$0.56_{0.01}$	$0.00_{0.00}$	$0.01_{0.00}$	$0.58_{0.02}$	$0.69_{0.01}$	$0.15_{0.02}$	$0.18_{0.01}$	$0.73_{0.01}$	$0.71_{0.01}$
RCO*	$0.00_{0.00}$	$0.01_{0.00}$	$0.22_{0.02}$	$0.56_{0.01}$	$\underline{0.04}_{0.00}$	$0.04_{0.00}$	$0.58_{0.02}$	$0.70_{0.02}$	$0.15_{0.02}$	$0.18_{0.02}$	$0.73_{0.01}$	$0.71_{0.00}$
SOM*	$\underline{0.72}_{0.03}$	$\underline{0.56}_{0.01}$	$0.22_{0.02}$	$0.58_{0.01}$	$0.01_{0.00}$	$\underline{0.07}_{0.00}$	$0.55_{0.02}$	$0.67_{0.01}$	$0.14_{0.01}$	$0.17_{0.01}$	$0.55_{0.01}$	$0.70_{0.01}$
HER	$0.01_{0.00}$	$0.02_{0.00}$	$0.22_{0.02}$	$0.56_{0.00}$	$0.00_{0.00}$	$0.07_{0.00}$	$0.57_{0.02}$	$0.69_{0.02}$	$0.18_{0.01}$	$0.15_{0.01}$	$0.75_{0.01}$	$0.74_{0.01}$
Ours	$\mathbf{0.86}_{0.03}$	$\mathbf{0.67}_{0.02}$	$0.23_{0.02}$	$0.54_{0.01}$	$\mathbf{0.42}_{0.00}$	$\mathbf{0.28}_{0.00}$	$0.48_{0.02}$	$0.53_{0.01}$	$\mathbf{0.96}_{0.01}$	$\mathbf{0.73}_{0.01}$	$0.78_{0.01}$	$0.67_{0.01}$

Method	Gpt-oss-20b				Qwen3.8-27B				Gemma-4-12B			
	ASR	JS	*MMLU*	*ARC-E*	ASR	JS	*MMLU*	*ARC-E*	ASR	JS	*MMLU*	*ARC-E*
Base	$0.01_{0.01}$	$0.01_{0.00}$	$0.28_{0.02}$	$0.36_{0.01}$	$0.02_{0.01}$	$0.08_{0.02}$	$0.51_{0.02}$	$0.84_{0.01}$	$0.07_{0.01}$	$0.10_{0.01}$	$0.73_{0.01}$	$0.72_{0.02}$
SD	$0.44_{0.02}$	$0.26_{0.01}$	$0.27_{0.03}$	$0.37_{0.01}$	$0.55_{0.02}$	$0.37_{0.02}$	$0.55_{0.01}$	$0.84_{0.01}$	$0.00_{0.00}$	$0.05_{0.00}$	$0.25_{0.02}$	$0.26_{0.02}$
RDO*	$0.02_{0.00}$	$0.01_{0.00}$	$0.27_{0.03}$	$0.38_{0.01}$	$0.22_{0.01}$	$0.19_{0.02}$	$0.50_{0.01}$	$0.83_{0.01}$	$0.00_{0.00}$	$0.05_{0.00}$	$0.26_{0.01}$	$0.24_{0.01}$
RCO*	$0.10_{0.01}$	$0.06_{0.01}$	$0.27_{0.02}$	$0.38_{0.02}$	$0.04_{0.01}$	$0.09_{0.01}$	$0.56_{0.02}$	$0.84_{0.01}$	$0.00_{0.00}$	$0.05_{0.00}$	$0.25_{0.02}$	$0.25_{0.01}$
SOM*	$\mathbf{0.94}_{0.01}$	$\mathbf{0.66}_{0.02}$	$0.28_{0.03}$	$0.38_{0.01}$	$\underline{0.65}_{0.03}$	$\underline{0.44}_{0.01}$	$0.59_{0.03}$	$0.84_{0.01}$	$\underline{0.89}_{0.01}$	$\underline{0.63}_{0.01}$	$0.63_{0.01}$	$0.65_{0.01}$
HER	$0.14_{0.01}$	$0.10_{0.02}$	$0.27_{0.03}$	$0.37_{0.01}$	$0.03_{0.01}$	$0.09_{0.02}$	$0.51_{0.02}$	$0.84_{0.02}$	$0.50_{0.02}$	$0.38_{0.01}$	$0.74_{0.01}$	$0.72_{0.01}$
Ours	$\underline{0.92}_{0.02}$	$\underline{0.62}_{0.01}$	$0.27_{0.02}$	$0.36_{0.02}$	$\mathbf{0.85}_{0.01}$	$\mathbf{0.59}_{0.01}$	$0.56_{0.02}$	$0.84_{0.01}$	$\mathbf{0.92}_{0.01}$	$\mathbf{0.66}_{0.01}$	$0.66_{0.01}$	$0.72_{0.02}$

Table 2: Ablation study for Algorithm 1. We report ASR at temperature 0 for variants obtained by removing individual components, namely the iterative procedure (Line 5), the refusal set update (Lines 6–8), and the adaptive prompt filtering (Lines 9–10).

Variant	Qwen2.5-3B	Llama-2-7B	Gemma-2-9B
Full method	**0.93**	**0.86**	**0.95**
w/o refusal-set update	0.87	0.27	0.31
w/o prompt filtering	0.77	0.14	0.55
w/o update and filtering	0.91	0.00	0.47
w/o iterations	0.84	0.00	0.19

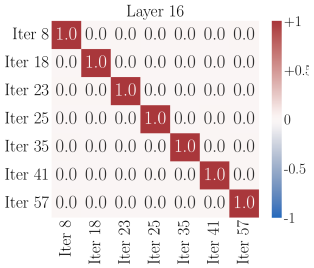


Figure 1: Cosine similarity of same-layer refusal directions for Llama-2-7B.

decreases by more than 0.04 relative to the original models. On Gemma-4-12B, *Ours* achieves comparable ASR to the strongest SOM configuration while retaining higher accuracy on both benchmarks. A larger trade-off emerges on Mistral-7B, where ASR increases from at most 0.04 among baselines to 0.42, but MMLU and ARC-E decrease by 0.09 and 0.16, respectively. This pattern may reflect greater overlap between refusal-related representations and those supporting general capabilities, making utility preservation more challenging under successive interventions. A more detailed Pareto analysis across ASR and both utility benchmarks, including model selections other than the validation-optimal one, is provided in Appendix A.2.

Ablation analysis. We analyze three design choices underlying our method, namely iterative steering, refusal-set updates, and adaptive prompt filtering, in Table 2, with extended results in Table 6. Their design reflects our central finding that refusal is *non-stationary* under steering. Whereas Arditi et al. (2024) estimate a single direction from the original model, our procedure re-estimates refusal and recomputes layer-specific directions after each intervention, adapting subsequent steps to the remaining refusal behavior under the benign-behavior constraint. The ablations support this adaptive design to handle the non-stationarity of refusal, with the full method improving ASR by up to 0.86 over the single-iteration variant. In particular, refusal-set updates and prompt filtering address distinct aspects of this behavior. On one hand, updating the refusal set tracks changes in refusal expression, yielding ASR gains of +0.06, +0.59, and +0.64 on Qwen2.5-3B, Llama-2-7B, and

Gemma-2-9B, respectively. For example, Table 8 shows that the dominant token beginning refusal responses on Gemma-2-9B changes from `I` to `Please`, `That`, and `You` across iterations, illustrating why a fixed refusal vocabulary can become less informative while refusal persists. On the other hand, adaptive prompt filtering retains prompts that remain informative for estimating the current refusal direction, yielding gains of $+0.16$, $+0.72$, and $+0.40$. The larger effects on Llama-2-7B and Gemma-2-9B highlight the importance of updating both refusal estimates and direction-estimation data when a single intervention is insufficient.

Mechanistic interpretability. Beyond uncensoring effectiveness, we analyze the ordered intervention sequence to examine where steering acts and how refusal-related directions change. The layer-selection distributions in Figure 2 and Appendix B show that interventions concentrate at particular layers for each model, some of which are selected repeatedly. For example, layer 16 of Llama-2-7B is selected seven times, indicating it offers the greatest reductions in refusal among candidate interventions satisfying the benign-behavior constraint. Moreover, directions selected at the *same layer* across iterations show near-zero cosine similarity (Figure 1), and other models show similar behavior in Appendix B. Successive interventions, therefore, target distinct directions rather than repeatedly removing the same one, showing that refusal can persist after a single direction is removed. As shown in Appendix B, directions from *different layers*, particularly neighboring ones, are initially similar but become less correlated as steering progresses. Lastly, Figure 2 shows PCA projections of harmful and benign prompt representations. At iteration 0, the distributions are clearly separated, consistent with Arditi et al. (2024). As steering progresses, they begin to overlap before the model fully stops refusing harmful prompts, indicating that reduced separation alone is insufficient to eliminate refusal. At later iterations, corresponding to higher ASR, only a small fraction of harmful prompts remains refused. In conclusion, our findings suggest that refusal involves multiple layer-specific components whose relevance changes during steering, supporting, alongside the behavioral observations and ablation results, its *non-stationarity* and motivating their iterative identification and removal.

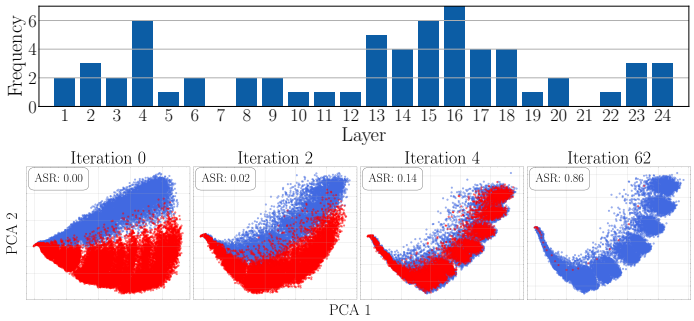


Figure 2: Layer-selection frequencies (top) and PCA projections of stacked layer activations for harmful (red) and benign (blue) prompts on Llama-2-7B at selected steering iterations (bottom).

6 Conclusion

In this work, we show the importance of accounting for the *non-stationarity* of refusal under steering. Our method treats refusal as a target that changes under intervention, updating the refusal vocabulary, filtering training prompts, and recomputing layer-specific directions at each step under a benign-behavior constraint, with ablations supporting the importance of these adaptive components. As a result, our approach achieves more effective uncensoring than the evaluated baselines, showing that model scale and stronger alignment do not necessarily ensure robustness against uncensoring. Beyond uncensoring effectiveness, the ordered intervention sequence supports a localized mechanistic interpretation of refusal. Rather than being fully characterized by a single stable direction, refusal appears to involve multiple evolving, layer-specific components whose relevance changes throughout steering. Our findings thus suggest that current alignment mechanisms rely on structured yet fragile internal representations, motivating more robust safety training and the use of mechanistic interpretability as a practical tool for auditing LLM safety. Future work should investigate whether these findings transfer to other architectures and modalities and develop stronger auditing methods and safety mechanisms for publicly available models whose weights can be directly modified.

AI USE STATEMENT

In this work, we used large language models exclusively for text polishing and minor exposition refinements. All substantive research content, methodology, and scientific conclusions were developed entirely by the authors. We reviewed all AI-assisted edits and take responsibility for the final content of this work.

ETHICS STATEMENT

Our work aims to improve understanding of refusal in LLMs and its robustness to direct model modification. Although the proposed method could be misused to bypass safeguards in open-weight models, examining these vulnerabilities is necessary to assess the protection offered by current alignment techniques. Our analysis reveals how refusal changes under successive interventions, helping identify weaknesses that fixed steering attacks may overlook. These findings can inform more thorough safety audits, strengthen robustness evaluations, and guide the development of alignment methods that better resist model uncensoring attacks.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide a detailed description of our experimental setup in Section 4, including datasets, models, and adversarial attacks, along with their exact sources in Appendix D. Furthermore, our source code has been included as part of the supplementary material and is available at this repository.

REFERENCES

M. Andriushchenko, F. Croce, and N. Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction. In *Neural Information Processing Systems*, 2024.

M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, and Others. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2024.

P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. 2017.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

K. Collier and J. Cui. Chatgpt safety systems can be bypassed to get weapons instructions. https://www.nbcnews.com/tech/security/chatgpt-safety-systems-can-bypassed-weapons-instructions-rcna225788, 2025.

J. Engels, I. Liao, E. J. Michaud, W. Gurnee, and M. Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

M. Gault. Poems can trick ai into helping you make a nuclear weapon. https://www.wired.com/story/poems-can-trick-ai-into-helping-you-make-a-nuclear-weapon/, 2025.

D. Glukhov, I. Shumailov, Y. Gal, N. Papernot, and V. Papyan. Position: Fundamental limitations of llm censorship necessitate new approaches. In *Forty-first International Conference on Machine Learning*, 2024.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and Others. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning ai with shared human values. *International Conference on Learning Representations*, 2021a.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021b.

H. Jin, L. Hu, X. Li, P. Zhang, C. Chen, J. Zhuang, and H. Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*, 2024.

H. Li, H. Gao, Z. Zhao, Z. Lin, J. Gao, and X. Li. Llms caught in the crossfire: Malware requests and jailbreak challenges. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Y. Lin, P. He, H. Xu, Y. Xing, M. Yamada, H. Liu, and J. Tang. Towards understanding jailbreak attacks in llms: A representation space analysis. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Z. Lin, Z. Li, X. Liao, and X. F. Wang. Consiglieres in the shadow: Understanding the use of uncensored large language models in cybercrimes. *arXiv preprint arXiv:2508.12622*, 2025.

M. Mazeika, L. Phan, X. Yin, A. Zou, and Others. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2023.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, and Others. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

W. Pan, Z. Liu, Q. Chen, X. Zhou, H. Yu, and X. Jia. The hidden dimensions of llm alignment: A multi-dimensional analysis of orthogonal safety directions. *arXiv preprint arXiv:2502.09674*, 2025.

K. Park, Y. Joong Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

G. Piras, R. Mura, F. Brau, L. Oneto, F. Roli, and B. Biggio. Som directions are better than one: Multi-directional refusal suppression in language models. In *AAAI Conference on Artificial Intelligence*, 2026.

PyTorch. Pytorch hub. https://pytorch.org/hub/, 2025.

X. Qi, Y. Zeng, T. Xie, P. Y. Chen, R. Jia, P. Mittal, and P. Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Neural information processing systems*, 2023.

M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874, 2024.

A. Robey, E. Wong, H. Hassani, and G. J. Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL https://arxiv.org/abs/2402.10260.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Lintang Sutawika, Leo Gao, Hailey Schoelkopf, Stella Biderman, Jonathan Tow, Baber Abbasi, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Aflah, and Others. Eleutherai/lm-evaluation-harness: Major refactor, December 2023. URL https://doi.org/10.5281/zenodo.10256836.

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

G. Team, M. Riviere, S. Pathak, P. G. Sessa, and Others. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, and Others. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, and Others. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, and Others. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

P. E. Weidmann. Heretic: Fully automatic censorship removal for language models. https://github.com/p-e-w/heretic, 2025.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and Others. Transformers: State-of-the-art natural language processing. In *Conference on empirical methods in natural language processing: system demonstrations*, 2020.

T. Wollschläger, J. Elstner, S. Geisler, V. Cohen-Addad, S. Günnemann, and J. Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

X. Wu, G. Hong, Y. Chen, M. X. Liu, F. Jin, X. Pan, J. Dai, and B. Liu. When bots take the bait: Exposing and mitigating the emerging social engineering attack in web automation agent. *arXiv preprint arXiv:2601.07263*, 2026.

T. Xie, X. Qi, Y. Zeng, Y. Huang, U. M. Sehwag, and Others. Sorry-bench: Systematically evaluating large language model safety refusal. *arXiv preprint arXiv:2406.14598*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, and Others. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, and Others. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

X. Yang, X. Wang, Q. Zhang, L. Petzold, W. Y. Wang, X. Zhao, and D. Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks. Improving alignment and robustness with circuit breakers. In *Neural Information Processing Systems*, 2024.