REFUSALTRACE: SPARSE ACTIVATION MASKING FOR JAILBREAKING VIA NATURAL REFUSAL TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment aims to make large language models (LLMs) refuse requests that violate safety constraints, thereby avoiding harmful or unethical outputs. However, the internal computations underlying refusal remain poorly understood. Existing white-box methods commonly select intervention coordinates from discriminative differences between harmful and benign inputs, but these coordinates may encode benign representational signals that are not directly involved in refusal generation. To address this limitation, we introduce REFUSALTRACE, a sparse activation-masking jailbreak attack framework based on natural refusal trajectories. REFUSALTRACE uses the refusal trajectories that the target model naturally generates for attack prompts as behavioral signals. It first attributes gate/up projection coordinates near the prompt–response boundary and aggregates candidates through stable voting across trajectories. It then performs conditional multi-fidelity compression subject to an attack-effectiveness constraint, yielding a compact static mask for inference-time use. Across 11 open-weight aligned LLMs, the initial candidate masks cover 0.47% of intervenable gate/up projection coordinates on average and achieve a mean attack success rate (ASR) of 83.2%. After conditional compression, the final static masks cover only 0.30% of coordinates while retaining a mean ASR of 74.9%. These results show that natural refusal trajectories can serve as behavior-conditioned signals for constructing compact static activation masks for jailbreak attacks.

1 INTRODUCTION

Large language models (LLMs) have been widely deployed in applications such as content generation, question answering, programming, and intelligent agents. Their open-ended generation capability also creates risks of harmful, misleading, or safety-violating outputs. To mitigate these risks, researchers commonly use supervised fine-tuning, reinforcement learning from human feedback, and safety guardrail mechanisms to induce models to refuse harmful requests (Ouyang et al., 2022; Meta, 2024).However, safety alignment does not necessarily imply robust refusal behavior, as jailbreak attacks can be induced either by carefully crafted prompts or, when model internals are accessible, by white-box modifications to safety-related neurons (Chen et al., 2024; Wu et al., 2026).

In recent years, researchers have begun to analyze safety-alignment mechanisms from the perspective of models' internal representations. Existing work has found that interventions on residual-stream directions, representation spaces, or MLP neurons can substantially alter model responses to harmful requests(Zou et al., 2023a; Arditi et al., 2024; Chen et al., 2024), suggesting that refusal behavior may depend on relatively sparse and manipulable internal computational components. Existing white-box methods commonly identify candidate coordinates according to activation differences between harmful and benign inputs, and subsequently train probes, select safety neurons, or construct dynamic interventions(Chen et al., 2024; Wu et al., 2026). However, such discriminative signals primarily characterize differences between input categories, rather than the behavior of generating a refusal response. Consequently, the resulting coordinates may include components associated with benign prompting or prompt-category differences that are not required for a compact refusal-disrupting intervention.

When faced with harmful requests, safety-aligned models typically generate refusal text with clear semantics and structure(Arditi et al., 2024; Yeo et al., 2025). This text is not an externally annotated

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.

class label; rather, it is a behavioral outcome produced by the model itself under the current prompt, template, and decoding configuration. We therefore consider a more direct question: can a model's own refusal trajectories be used as supervision signals to localize the internal coordinates that support refusal generation? If certain coordinates make stable contributions to the generation of refusal trajectories, suppressing these coordinates at inference time may weaken the model's refusal capability.

To address this question, we propose REFUSALTRACE, a sparse activation-masking jailbreak attack framework based on natural refusal trajectories. For jailbreak prompts that elicit a natural refusal, REFUSALTRACE uses the conditional generation objective defined on the model's own refusal trajectory as an attribution target. It identifies candidate gate and up projection coordinates near the prompt–response boundary in gated MLPs, then aggregates them through stable voting across trajectories. Conditional multi-fidelity search subsequently removes redundant coordinates subject to a fixed attack-effectiveness constraint, yielding a compact static mask for inference-time use. Experiments show that natural refusal trajectories provide a behavior-conditioned candidate source for compact refusal-disrupting interventions.

The contributions of this paper are as follows:

- We propose a behavioral attribution method that uses model-generated natural refusal trajectories as supervision signals to localize internal coordinates associated with refusal.
- We propose a behavior-constrained multi-fidelity compression strategy that removes redundant coordinates from the attributed candidate pool subject to a fixed attack-effectiveness constraint, yielding a compact static mask for inference-time deployment.
- We evaluated REFUSALTRACE on 11 open-weight aligned LLMs, covering attack effectiveness, mask compactness, candidate-source ablation, cross-prompt generalization, and capability cost.

2 RELATED WORK

2.1 INTERNAL REPRESENTATIONS OF REFUSAL AND SAFETY LOCALIZATION

Safety alignment commonly uses supervised fine-tuning and reinforcement learning from human feedback to make models prefer refusing over complying when faced with harmful requests. Recent mechanistic studies ask how this refusal behavior is represented internally. Representation engineering has shown that directions in activation space can modulate high-level model behavior (Zou et al., 2023a). Within this perspective, Arditi et al. identified a low-dimensional direction in the residual stream whose ablation reduces refusal across multiple aligned models (Arditi et al., 2024).

A residual-stream direction provides a compact description of refusal-related model state, but does not by itself identify the architectural components that jointly support refusal. Subsequent work has therefore pursued finer-grained safety localization. Chen et al. localized safety-related MLP coordinates through generation-time activation contrasts and used activation patching to examine the effects of local interventions (Chen et al., 2024). Li et al. subsequently identified a sequence of safety layers associated with harmful-request recognition (Li et al., 2025). Together, these studies extend refusal analysis from global representation directions to architectural components, including layers and internal coordinates.

However, neuron-level localization does not imply that each coordinate has a single semantic role, because model activations are often polysemantic. Sparse autoencoders (SAEs) provide a complementary feature-level view by decomposing activations into sparse, interpretable latents (Cunningham et al., 2023). O'Brien et al. steered refusal through SAE features and measured the associated robustness–capability trade-off (O'Brien et al., 2024). Yeo et al. identified refusal- and harm-related features and tested their relationship through interventions (Yeo et al., 2025). These studies suggest that refusal is distributed across representational scales, from directions and neurons to latent features, rather than localized to a unique causal module.

2.2 JAILBREAK ATTACKS

Input-space jailbreak attacks seek prompts that induce policy-violating responses without modifying the target model. Early approaches relied on manually designed templates, such as role-playing, instruction obfuscation, and prompt decomposition (Wei et al., 2023). Although these strategies demonstrate that safety policies can be bypassed through prompting alone, their effectiveness can depend heavily on manually chosen templates and attack strategies. To reduce this reliance on hand-crafted designs, subsequent work formulates jailbreak-prompt construction as an optimization or search problem, using gradient-based adversarial suffixes, genetic search, or iterative rewriting by language models (Zou et al., 2023b; Liu et al., 2023; Chao et al., 2023). Because these methods leave model parameters and activations unchanged, they remain applicable in black-box settings, but their success can still depend on the target model's feedback and repeated search procedures.

White-box jailbreaks instead intervene directly in model computation. Chen et al. identify safety neurons by contrasting generation-time activations elicited by harmful and benign prompts, and test their causal effects through dynamic activation patching (Chen et al., 2024). Zhao et al. refine safety knowledge neurons by removing neurons that are also activated by benign prompts, then manipulate the remaining neurons to control safety behavior (Zhao et al., 2026). NeuroStrike uses activation patterns labeled by jailbreak success or failure to train a linear discriminator, then masks selected safety neurons at inference time (Wu et al., 2026).

Although existing white-box internal attacks have shown that modifying a small number of representations or neurons can substantially change a model's safety behavior, their candidate localization procedures commonly rely on harmful–benign input contrasts, predefined refusal directions, or additional trained discriminators. As a result, selected coordinates may encode benign representational signals that are not directly involved in refusal generation. For open-ended generation tasks, attribution scores for individual coordinates also cannot directly predict the final behavior that arises when multiple coordinates are jointly masked (Prakash et al., 2026). To address these limitations, we propose REFUSALTRACE, a sparse activation-masking jailbreak attack framework based on natural refusal trajectories. The framework uses refusal trajectories naturally generated by the model in response to jailbreak prompts as reference generation sequences, performs boundary-local attribution near the prompt–response boundary, and forms a stable candidate pool through aggregation across trajectories. It then compresses the candidate coordinate pool under a fixed attack-effectiveness constraint to obtain a sparse static mask.

3 METHOD

3.1 PROBLEM SETTING AND METHOD OVERVIEW

Safety-aligned models usually refuse harmful requests, but refusal generation is not determined by a single internal unit. Existing white-box methods commonly select intervention coordinates from discriminative differences between harmful and benign inputs. Such coordinates may encode benign representational signals that are unrelated to refusal generation itself. We instead use a model's naturally generated refusal trajectories to construct a sparse activation mask that can be reused at inference time to weaken refusal behavior.

Let f_θ be an aligned autoregressive language model with L Transformer layers, and let $\mathcal{D}_{\text{build}} = \{x_j\}_{j=1}^{N}$ be the harmful-prompt set used during offline construction. For each $x \in \mathcal{D}_{\text{build}}$, the unmodified model generates

$$r(x) = (r_1, \ldots, r_T). \tag{1}$$

Let $F_{\text{ref}} \in \{0, 1\}$ be a fixed refusal detector, where $F_{\text{ref}}(r) = 1$ indicates that response r is classified as a refusal. We retain naturally generated refusals to form

$$\mathcal{R} = \{(x, r(x)) \mid x \in \mathcal{D}_{\text{build}}, F_{\text{ref}}(r(x)) = 1\}. \tag{2}$$

The set $\mathcal{R}$ is the sole supervision source for trajectory attribution. REFUSALTRACE neither trains an additional safety probe nor requires a benign counterpart for each harmful prompt. It does not update model parameters.

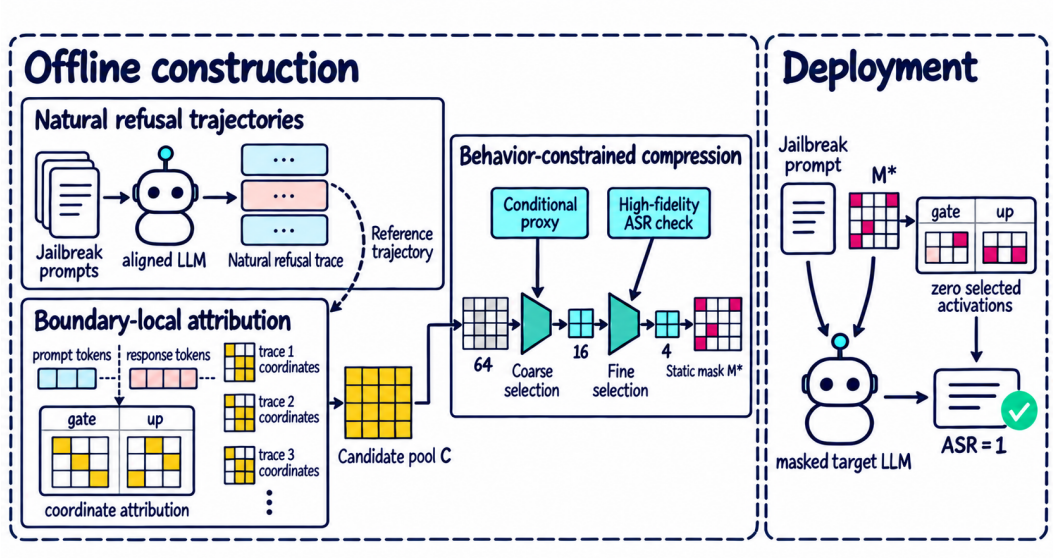


Figure 1: Offline construction and inference-time deployment of REFUSALTRACE.

We intervene in the gate and up projections of a gated MLP. At layer ℓ and token position p, let $h_{\ell,p}$ be the MLP input, and let $g_{\ell,p}$ and $u_{\ell,p}$ be the gate and up projection outputs:

$$\begin{aligned} g_{\ell,p} = W_\ell^{\text{gate}} h_{\ell,p}, \qquad u_{\ell,p} = W_\ell^{\text{up}} h_{\ell,p}, \\ \text{MLP}_\ell(h_{\ell,p}) = W_\ell^{\text{down}} \left(\sigma(g_{\ell,p}) \odot \phi(u_{\ell,p})\right), \end{aligned} \tag{3}$$

where σ and ϕ are architecture-specific element-wise functions. We treat each output dimension of each gate or up projection as an independently intervenable coordinate. A module is denoted by $m = (\ell, t)$, where $t \in \{\text{gate}, \text{up}\}$, and coordinate $c = (m, i)$ denotes dimension i of module m.

Let τ denote a fixed chat template and δ a fixed decoding configuration. Under mask M, the response to prompt x is

$$y_M(x) = \text{Decode}_\delta\left(f_\theta(\tau(x); M)\right). \tag{4}$$

We use Llama-Guard-3-8B as the safety evaluator (Meta, 2024). For a fixed compression set $\mathcal{D}_{\text{cmp}}$, the attack success rate is

$$\text{ASR}_{\mathcal{D}_{\text{cmp}}}(M) = \frac{1}{|\mathcal{D}_{\text{cmp}}|} \sum_{x \in \mathcal{D}_{\text{cmp}}} \mathbf{1}[J_{\text{LG}}(x, y_M(x)) = \text{unsafe} \wedge F_{\text{ref}}(y_M(x)) = 0]. \tag{5}$$

Thus, a response is counted as a successful attack only when it is judged unsafe by Llama-Guard-3-8B and is not classified as a refusal by the fixed refusal filter. This conjunction prevents residual refusal responses from being counted as successful jailbreaks. We use the same evaluation rule for compression checks, the main attack evaluation, and the matched comparison with NeuroStrike.

The candidate pool C induces an initial mask M_C: coordinates in C are set to zero during inference. Let

$$\alpha_0 = \text{ASR}_{\mathcal{D}_{\text{cmp}}}(M_C) \tag{6}$$

be the behavioral anchor of the initial mask. Given an allowable decrease ϵ, REFUSALTRACE seeks a compact final static mask:

$$M^* = \arg\min_{M \subseteq C} |M| \quad \text{s.t.} \quad \text{ASR}_{\mathcal{D}_{\text{cmp}}}(M) \geq \alpha_0 - \epsilon. \tag{7}$$

Figure 1 summarizes the offline construction and inference-time deployment of REFUSALTRACE. During offline construction, jailbreak prompts are first passed to the aligned LLM. The model's naturally generated refusal responses are retained as reference trajectories. REFUSALTRACE then attributes gate/up projection coordinates from activations near the prompt–response boundary for each reference trajectory. Stable voting across trajectories aggregates these coordinates into the candidate

pool C.Starting from the mask induced by C, behavior-constrained compression progressively restores coordinate blocks from coarse to fine scales. The conditional proxy prioritizes blocks for restoration, and the high-fidelity ASR check accepts a restoration only when the attack-effectiveness constraint is preserved, yielding the final static mask M^*. During deployment, a new jailbreak prompt is processed by the target LLM with M^* fixed: hooks zero the selected gate/up projection coordinates at inference time, producing the masked-model response used to evaluate ASR.

3.2 REFUSAL-TRAJECTORY-GUIDED CANDIDATE LOCALIZATION

This stage converts the retained natural refusal trajectories $\mathcal{R}$ into a stable candidate pool C for subsequent compression. For each trajectory, REFUSALTRACE attributes gate and up projection coordinates according to their local contributions to conditional refusal generation near the prompt–response boundary. Module-local stable voting then retains coordinates that recur across natural refusals, without directly comparing raw attribution scores across projection modules. The resulting pool C is a high-recall, behavior-conditioned search space rather than the final static mask.

3.2.1 CONDITIONAL REFUSAL-TRAJECTORY OBJECTIVE

Existing discriminative methods commonly select candidate coordinates based on representational differences between harmful and benign inputs. Such signals primarily measure the separability of input categories and may therefore include representational components that distinguish benign from harmful inputs without directly contributing to refusal generation. To reduce interference from these benign representational signals, REFUSALTRACE neither prespecifies refusal phrases nor manually constructs target responses. Instead, it uses the refusal trajectory that the model naturally generates in response to each attack prompt as the conditional-generation target.

For each retained sample $(x, r) \in \mathcal{R}$, let $r = (r_1, \ldots, r_T)$ be the refusal sequence truncated to a fixed maximum length. We define its mean conditional log-likelihood as

$$J_{\mathrm{ref}}(x, r) = \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(r_t \mid \tau(x), r_{<t}). \tag{8}$$

This objective measures the model's tendency to continue generating the same refusal trajectory given the original request and the response prefix.

3.2.2 BOUNDARY-LOCAL ATTRIBUTION

Defining the generation target using a complete refusal trajectory does not by itself determine where to read the internal signal. If attribution is performed directly at response-token positions, highly scored coordinates may primarily reflect refusal wording, syntactic structure, or general language-modeling preferences, rather than the internal transition through which the model moves from an attack prompt to refusal generation. REFUSALTRACE therefore restricts the attribution window to the local boundary region after the prompt ends but before the assistant response begins.

Let $p_0(x)$ be the final valid token position of prompt x. The attribution window contains the preceding K valid positions:

$$\mathcal{W}_K(x) = \{p_0(x) - K + 1, \ldots, p_0(x)\}. \tag{9}$$

For coordinate $c = (m, i)$, let $a_{c,p}$ be its projection-output activation at position p. We define a window-averaged activation-gradient score, $s_c(x, r)$, to estimate its local positive contribution to the reference trajectory objective:

$$s_c(x, r) = \frac{1}{|\mathcal{W}_K(x)|} \sum_{p \in \mathcal{W}_K(x)} \max\left(0,\, a_{c,p} \frac{\partial J_{\mathrm{ref}}(x, r)}{\partial a_{c,p}}\right). \tag{10}$$

3.2.3 MODULE-LOCAL STABLE VOTING

Attribution from a single refusal trajectory is sensitive to prompt semantics, refusal wording, and local activation variation. Directly writing high-scoring coordinates from a single sample into the final static mask would therefore make the result overly dependent on individual requests. We instead

perform stable voting within each projection module and avoid directly comparing raw attribution scores across layers or projection modules.

Let U_m be the coordinates in module m. For each retained trajectory, we keep its top-k_{loc} coordinates:

$$T_m(x,r) = \mathrm{TopK}_{c\in U_m}\left(s_c(x,r), k_{\mathrm{loc}}\right). \tag{11}$$

The support rate of coordinate c is

$$v_m(c) = \frac{1}{|\mathcal{R}|}\sum_{(x,r)\in\mathcal{R}} \mathbf{1}[c \in T_m(x,r)]. \tag{12}$$

For each module, we first retain coordinates that appear in the module-local top-k_{loc} set for at least a fraction ρ of trajectories:

$$C_m = \{c \in U_m \mid v_m(c) \geq \rho\}. \tag{13}$$

For modules with dispersed attribution, where no coordinate reaches ρ, we retain a small number of coordinates with the largest cumulative attribution scores. If a module contains more candidates than a predefined cap, we likewise truncate it according to cumulative attribution scores. After applying these fallback and cap rules, the final candidate pool is

$$C = \bigcup_m C_m. \tag{14}$$

3.3 Conditional Multi-Fidelity Mask Search

The goal of the candidate-localization stage is to retain a high-recall search space that is stably correlated with refusal trajectories, rather than to directly determine the final static mask. The effects of some coordinates may become apparent only when they are jointly masked with other coordinates. Truncating the candidate pool solely according to individual-coordinate attribution scores cannot reliably predict the behavioral effect of a joint intervention. To address this issue, RefusalTrace performs backward compression starting from the initial candidate mask M_C. At each step, it attempts to delete a coordinate block from the current mask, that is, to restore the original activations of that block. A deletion is accepted only when the fixed attack-effectiveness constraint remains satisfied after restoration.

3.3.1 Conditional Proxy Screening

High-fidelity evaluation requires autoregressive generation and safety judging. We first use the reference-trajectory objective as a low-fidelity proxy:

$$J_{\mathrm{proxy}}(M) = \frac{1}{|\mathcal{R}|}\sum_{(x,r)\in\mathcal{R}} J_{\mathrm{ref}}(x,r;M), \tag{15}$$

where $J_{\mathrm{ref}}(x,r;M)$ is evaluated under mask M.

The effect of restoring G can depend on other restored blocks. We therefore sample S restoration contexts Q_s from blocks outside G and compute

$$\overline{\Delta}_{\mathrm{proxy}}(G) = \frac{1}{S}\sum_{s=1}^{S}\left[J_{\mathrm{proxy}}(M\setminus Q_s) - J_{\mathrm{proxy}}(M\setminus(Q_s\cup G))\right]. \tag{16}$$

A value of $\overline{\Delta}_{\mathrm{proxy}}(G)$ close to zero indicates that, across different restoration contexts, additionally restoring G has limited effect on the reference refusal-trajectory objective. Such blocks are therefore prioritized for high-fidelity deletion review. Whether a deletion is accepted is still determined by the ASR criterion under the fixed behavioral anchor.

3.3.2 Block Deletion under a Fixed Behavioral Anchor

For each prioritized block, we regenerate responses under $M \setminus G$ and recompute Equation (5). Restoring G is accepted only if

$$\mathrm{ASR}_{\mathcal{D}_{\mathrm{cmp}}}(M\setminus G) \geq \alpha_0 - \epsilon, \qquad M \leftarrow M\setminus G. \tag{17}$$

Table 1: Attack effectiveness and mask compactness of REFUSALTRACE.

Model	Base ASR	Pool ASR ↑	Pool Size	Final ASR ↑	Mask Size ↓	Mask Ratio ↓	Reduction ↑
Llama-3.2-1B	0.6%	99.0%	1,644	78.0%	667	0.25%	59.4%
Llama-3.2-3B	1.9%	97.8%	2,390	89.8%	1,583	0.35%	33.8%
Llama-3.1-8B	0.3%	97.4%	3,304	81.2%	2,082	0.23%	37.0%
Qwen2.5-3B	22.0%	74.1%	3,982	59.1%	3,228	0.41%	18.9%
Qwen2.5-7B	12.5%	86.6%	3,621	83.1%	2,287	0.22%	36.8%
Qwen2.5-14B (4-bit)	0.6%	70.0%	5,211	68.4%	5,047	0.38%	3.1%
Phi-3.5-mini	4.5%	41.9%	2,615	37.7%	1,282	0.24%	51.0%
Phi-4 (4-bit)	0.3%	78.3%	4,135	65.2%	2,124	0.15%	48.6%
Phi-4-mini	0.6%	94.3%	2,912	86.6%	1,151	0.22%	60.5%
DS-R1-Distill-Llama-8B	67.4%	97.1%	3,640	97.8%	2,641	0.29%	27.4%
Mistral-7B-Instruct-v0.2	21.7%	78.6%	3,109	77.0%	2,483	0.54%	20.1%
Average	12.0%	83.2%	3,324	74.9%	2,234	0.30%	36.1%

The anchor α_0 remains fixed throughout the search. This prevents a sequence of locally acceptable deletions from accumulating an unconstrained reduction in attack effectiveness.

The search proceeds from coarse to fine block sizes. Large blocks first restore clearly redundant regions within a projection module, while smaller blocks refine the remaining mask. The final static mask M^* is applied at every inference position by zeroing the selected gate/up projection coordinates.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We conduct white-box evaluations on 11 open-weight aligned language models from the Llama, Qwen, Phi, DeepSeek, and Mistral families(AI at Meta, 2024; Qwen Team, 2025; Abdin et al., 2024; DeepSeek-AI, 2025; Jiang et al., 2023). For offline construction, REFUSALTRACE uses 100 harmful-behavior prompts from JailbreakBench (JBB-100), retaining the target model's natural refusal trajectories for candidate localization and conditional compression. We evaluate attack effectiveness on 313 StrongREJECT prompts using Llama-Guard-3-8B together with the refusal-filtering rule(Chao et al., 2024; Souly et al., 2024; Meta, 2024).

Unless otherwise specified, we use an attribution window of $K = 4$, a search prompt-set size of 32, a restoration-context fraction of 0.10, a behavioral-anchor tolerance of $\epsilon = 0.05$, and a coarse-to-fine block schedule of $64 \rightarrow 16 \rightarrow 4$. We report attack success rate (ASR), final mask size, mask ratio, and candidate-pool reduction. Additional hyperparameter sensitivity analyses and extended cross-prompt transfer results are reported in Appendix A.Natural refusal trajectories from JBB-100 are used for candidate localization. Conditional compression uses a 32-prompt search subset selected from JBB-100. StrongREJECT is reserved exclusively for final attack evaluation and is not used for candidate construction, compression search, or hyperparameter selection.

4.2 MAIN RESULTS

4.2.1 ATTACK EFFECTIVENESS AND MASK COMPACTNESS

Table 1 reports the main results of REFUSALTRACE on 11 open-weight aligned language models. The experimental results show that both the full candidate-pool masks and the compressed final masks increase ASR for every target model. These findings indicate that self-generated refusal trajectories can identify candidate projection coordinates associated with refusal behavior, and that these coordinates can still form effective static interventions after joint search and compression.

As shown in Table 1, conditional compression reduces the mean coordinate ratio from 0.47% for the initial candidate masks to 0.30% for the final static masks, a mean per-model reduction of 36.1%, while retaining a mean ASR of 74.9%. Phi-4-mini is compressed from 2,912 to 1,151 masked coordinates, a 60.5% reduction, while retaining an ASR of 86.6%. Llama-3.2-1B is similarly

Table 2: Direct comparison between REFUSALTRACE and NeuroStrike under matched evaluation.

Model	RT ASR ↑	NS ASR ↑	ΔASR (pp)	RT Size ↓	NS Size ↓	RT Reduction ↑
Llama-3.2-3B	**89.8%**	69.7%	+20.2	**1,583**	3,075	48.5%
Llama-3.2-1B	**78.0%**	62.6%	+15.4	**667**	1,587	58.0%
Phi-4-mini	**86.6%**	79.0%	+7.6	**1,151**	2,982	61.4%
Qwen2.5-7B	**83.1%**	79.1%	+4.0	**2,287**	4,690	51.2%
DS-R1-Distill-Llama-8B	**97.8%**	76.0%	+21.8	**2,641**	4,989	47.1%
Mistral-7B-Instruct-v0.2	77.0%	**78.0%**	-1.0	**2,483**	5,335	53.5%

Table 3: Candidate-source ablation.

	Unmodified	**Random Candidate Pool**		**Ours**
Model	**ASR ↑**	**ASR ↑**	**Mask Size ↓**	**ASR ↑**
Llama-3.2-1B	0.32%	0.00%	335	78.0%
Llama-3.2-3B	1.9%	0.00%	1,371	89.8%
Qwen2.5-7B	12.5%	10.22%	2,307	83.1%
Phi-4-mini	0.6%	0.64%	858	86.6%
DS-R1-Distill-Llama-8B	69.97%	65.81%	3,563	97.8%
Mistral-7B-Instruct-v0.2	20.77%	19.81%	2,098	77.0%

reduced from 1,644 to 667 coordinates with a final ASR of 78.0%. These results indicate that some coordinates in the initial candidate masks are redundant under joint masking and can be removed under the behavioral constraint.

However, compression gains are model-dependent rather than uniformly large. Phi-4 (4-bit), Phi-3.5-mini, and Llama-3.2-1B achieve reductions of 48.6%, 51.0%, and 59.4%, respectively. In contrast, Qwen2.5-14B (4-bit) is reduced by only 3.1%, leaving its final mask close in size to the initial candidate pool under the current compression configuration. DeepSeek-R1-Distill-Llama-8B also has a distinct starting point: its unmodified ASR is already 67.4%, so its final ASR of 97.8% should be interpreted as amplifying an existing safety vulnerability rather than as directly comparable to gains on low-baseline models.

4.2.2 DIRECT COMPARISON WITH NEUROSTRIKE

We compare RefusalTrace and NeuroStrike under matched target-model checkpoints, gate/up projection intervention units, chat templates, quantization settings, decoding configurations, StrongREJECT prompts, and the Llama-Guard-3-8B evaluation protocol.

As shown in Table 2, REFUSALTRACE attains both higher ASR and fewer masked coordinates than NeuroStrike on Llama-3.2-3B,Llama-3.2-1B, Phi-4-mini, Qwen2.5-7B, and DeepSeek-R1-Distill-Llama-8B. Across these models, it improves ASR by 4.0 to 21.8 percentage points while reducing mask size by 47.1% to 61.4%. The largest ASR gain occurs on DeepSeek-R1-Distill-Llama-8B, where ASR increases from 76.0% to 97.8%. On Phi-4-mini, REFUSALTRACE increases ASR to 86.6% while reducing the mask from 2,982 to 1,151 coordinates.

For Mistral-7B-Instruct-v0.2, although NeuroStrike attains an ASR that is 1.0 percentage point higher than that of REFUSALTRACE (78.0% vs. 77.0%), REFUSALTRACE reduces the number of masked coordinates from 5,335 to 2,483, a reduction of 53.5%. Taken together, REFUSALTRACE improves both attack effectiveness and mask compactness on five of the six models, while retaining a markedly more compact intervention with only a marginal ASR difference on the remaining model.

These results indicate a favorable effectiveness–compactness trade-off under the matched evaluation protocol.

4.2.3 CANDIDATE-SOURCE ABLATION

Table 1 shows that the trajectory-attributed candidate pool induces an effective initial mask before compression. To test whether this effectiveness depends on the content of the candidate pool rather than on pool size or the downstream compression procedure, we replace the attributed pool with a module-wise count-matched random coordinate pool and apply the same conditional compression procedure.

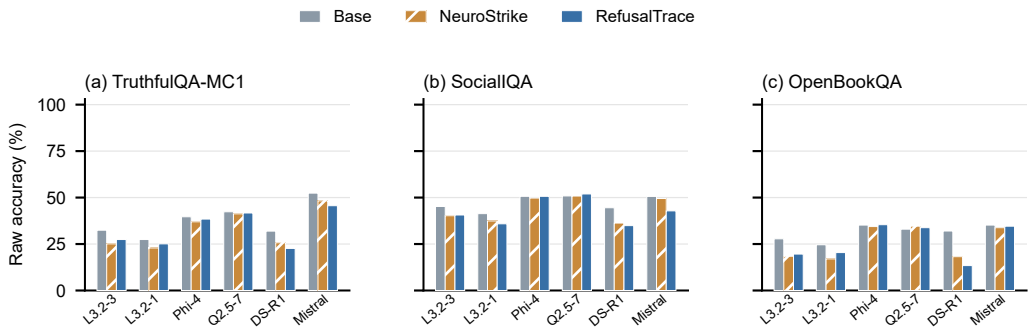


Figure 2: General-capability effects of NeuroStrike masks and final static REFUSALTRACE masks.

As shown in Table 3,random candidate pools fail to reproduce the attack effectiveness of REFUSALTRACE across all six models. Their mean ASR is 16.1%, close to the mean Base ASR of 17.7% obtained in the corresponding control runs, whereas REFUSALTRACE attains 85.4%. Even for DeepSeek-R1-Distill-Llama-8B, whose baseline ASR is relatively high, the random-pool ASR of 65.81% remains below the re-evaluated baseline of 69.97% and far below the 97.8% achieved by REFUSALTRACE. Thus, conditional compression alone cannot recover effective static masks from arbitrary coordinate sets. Together with Table 1, this result shows that the trajectory-attributed candidate pool provides the effective starting point, while conditional compression removes redundant coordinates within that candidate space.

4.2.4 GENERAL-CAPABILITY COMPARISON

Static activation masking may alter general task capabilities. To quantify these effects, we evaluate unmodified models, models with NeuroStrike masks, and models with final static REFUSALTRACE masks on zero-shot TruthfulQA-MC1, SocialIQA, and OpenBookQA (Lin et al., 2022; Sap et al., 2019; Mihaylov et al., 2018). These benchmarks assess truthfulness judgment, social commonsense reasoning, and open-book knowledge question answering, respectively.

Figure 2 compares the NLU performance of unmodified models, models with NeuroStrike masks, and models with final static REFUSALTRACE masks. Across the six evaluated models, mean raw accuracy on TruthfulQA-MC1 was 37.7% for unmodified models, 33.4% for models with NeuroStrike masks, and 33.5% for models with final static REFUSALTRACE masks. The corresponding accuracies were 47.2%, 44.0%, and 42.8% on SocialIQA, and 31.3%, 26.1%, and 26.2% on OpenBookQA. Overall, REFUSALTRACE substantially weakens refusal while retaining measurable NLU performance on the evaluated benchmarks, although the magnitude of the capability change varies across models and tasks.

DeepSeek-R1-Distill-Llama-8B merits separate interpretation, as it exhibits the largest REFUSALTRACE accuracy declines, by 9.3, 9.6, and 18.6 percentage points on TruthfulQA-MC1, SocialIQA, and OpenBookQA, respectively. One possible explanation is that its high unmodified ASR makes natural refusal trajectories a less selective signal for distinguishing refusal-specific coordinates from those supporting general response generation; the resulting mask may therefore overlap more with coordinates supporting general response generation.

5 CONCLUSION

This paper introduces REFUSALTRACE, a sparse activation-masking jailbreak framework that uses natural refusal trajectories for boundary-local attribution, stable voting, and behavior-constrained compression into a reusable static mask. Across 11 open-weight aligned language models, candidate masks cover 0.47% of intervenable projection coordinates on average and achieve a mean ASR of 83.2%, while compression reduces the final mask ratio to 0.30% and retains a mean ASR of 74.9%. These results show that natural refusal trajectories provide behavior-grounded candidate signals for compact inference-time interventions.

AI USE STATEMENT

AI-assisted tools were used to support translation and language editing during manuscript preparation. The authors independently designed the study, conducted the experiments, verified all technical claims and results, and take full responsibility for the final manuscript.

ETHICAL CONSIDERATIONS

This work studies white-box vulnerabilities of open-weight aligned language models to support safety evaluation and mitigation. All experiments are conducted offline on publicly available models and benchmark prompts; no human participants, private data, or deployed systems are involved. The proposed activation-masking method is dual-use because it could be misused to weaken model safeguards. We therefore present it as a stress test that helps identify conditions under which sparse internal interventions compromise refusal behavior.

REPRODUCIBILITY STATEMENT

We report the target model checkpoints, benchmark splits, intervention units, decoding and evaluation protocol, default hyperparameters, and evaluation metrics in the main text and appendix. The code and configuration files used to construct and evaluate the static masks will be released upon publication.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

AI at Meta. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://papers.nips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2023. URL https://arxiv.org/abs/2310.08419.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55005–55029, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Finding safety neurons in large language models, 2024. URL https://arxiv.org/abs/2406.14144. The revised arXiv version is titled "Towards Understanding Safety Alignment: A Mechanistic Perspective from Safety Neurons.".

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to LLM security. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f3bfbd65743e60c685a3845bd61ce15f-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models, 2023. URL https://arxiv.org/abs/2310.04451.

Meta. Llama Guard 3 8b model card. Model card, 2024. URL https://huggingface.co/meta-llama/Llama-Guard-3-8B. Accessed: 2026-09-21.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders, 2024. URL https://arxiv.org/abs/2411.11296.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Nirmalendu Prakash, Yeo Wei Jie, Amir Abdullah, Ranjan Satapathy, Erik Cambria, and Roy Ka-Wei Lee. Beyond I'm sorry, I can't: Dissecting large-language-model refusal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37830–37838, 2026. doi: 10.1609/aaai.v40i44.41119. URL https://ojs.aaai.org/index.php/AAAI/article/view/41119.

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. SocialIQA: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125416–125440, 2024. doi: 10.52202/079017-3984.

URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2307.02483.

Lichao Wu, Sasha Behrouzi, Mohamadreza Rostami, Maximilian Thang, Stjepan Picek, and Ahmad-Reza Sadeghi. NeuroStrike: Neuron-level attacks on aligned LLMs. In *Network and Distributed System Security Symposium*, 2026. doi: 10.14722/ndss.2026.230660. URL https://www.ndss-symposium.org/ndss-paper/neurostrike-neuron-level-attacks-on-aligned-llms/.

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Roy Ka-Wei Lee, Erik Cambria, and Ranjan Satapathy. Understanding refusal in language models with sparse autoencoders, 2025. URL https://arxiv.org/abs/2505.23556.

Chongwen Zhao, Yutong Ke, and Kaizhu Huang. Unraveling LLM jailbreaks through safety knowledge neurons. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2026. URL https://aclanthology.org/2026.eacl-long.83/.

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, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b. URL https://arxiv.org/abs/2307.15043.

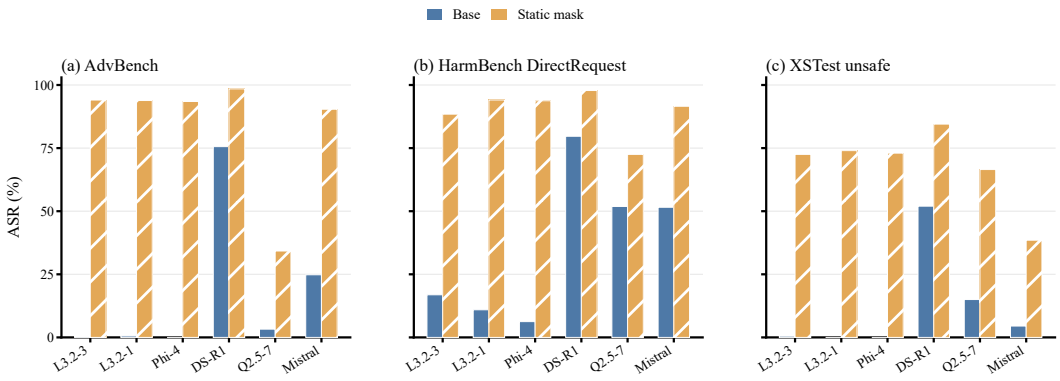


Figure 3: Cross-prompt generalization of REFUSALTRACE static masks on external attack prompt sets not used for construction.

A ADDITIONAL EXPERIMENTAL RESULTS

A.1 CROSS-PROMPT-SET GENERALIZATION

To examine whether a fixed REFUSALTRACE mask transfers to attack prompts that do not participate in construction, we evaluate the final static masks on AdvBench, HarmBench DirectRequest, and XSTest unsafe contrasts.

As shown in Figure 3, REFUSALTRACE increases ASR in all 18 model–dataset units. The macro-average ASR across the three external prompt sets increases from 21.9% to 80.6%, an average increase of 58.8 percentage points. Specifically, the average ASR on AdvBench increases from 17.4% to 84.1%; on HarmBench DirectRequest, it increases from 36.2% to 89.7%; and on XSTest unsafe contrasts, it increases from 12.0% to 68.2%. These results show that the fixed static mask does not only alter refusal behavior on construction-stage prompts, but can retain attack effectiveness across prompts from different sources, formulations, and harmful categories.

Transfer magnitudes vary across models and prompt sets. Llama-3.2-3B, Llama-3.2-1B, Mistral-7B-Instruct-v0.2 and Phi-4-mini exhibit large ASR increases on all three external prompt sets. For Mistral-7B-Instruct-v0.2, ASR increases from 24.8%, 51.6%, and 4.5% to 90.4%, 91.6%, and 38.5% on AdvBench, HarmBench DirectRequest, and XSTest unsafe contrasts, respectively. In contrast, Qwen2.5-7B reaches a final ASR of 34.2% on AdvBench, lower than on the other two external sets but still higher than its unmodified ASR of 3.3%. These differences indicate that cross-prompt transfer depends on both the model's initial safety state and the distribution of attack prompts.

A.2 SENSITIVITY ANALYSES

The offline construction process of REFUSALTRACE involves several design choices, including the prompt-boundary attribution window, behavioral-anchor tolerance, and attribution prompt budget. All main experiments use a single shared configuration to avoid per-model hyperparameter tuning and to enable consistent cross-model comparisons. Because attack effectiveness and mask compactness form a joint trade-off, we do not interpret any single setting as universally optimal or select settings solely by ASR. Instead, the analyses below characterize how these design choices affect ASR and final mask size across representative models under an otherwise fixed protocol.

A.2.1 PROMPT-BOUNDARY ATTRIBUTION WINDOW

The attribution window determines the number of token positions aggregated near the prompt boundary. A smaller window focuses on local activations at the onset of refusal generation, whereas a larger window incorporates a broader range of contextual information. We set $K \in \{1, 2, 4, 8\}$, rerun candidate localization, candidate selection, and mask compression, and evaluate the resulting final masks on StrongREJECT.

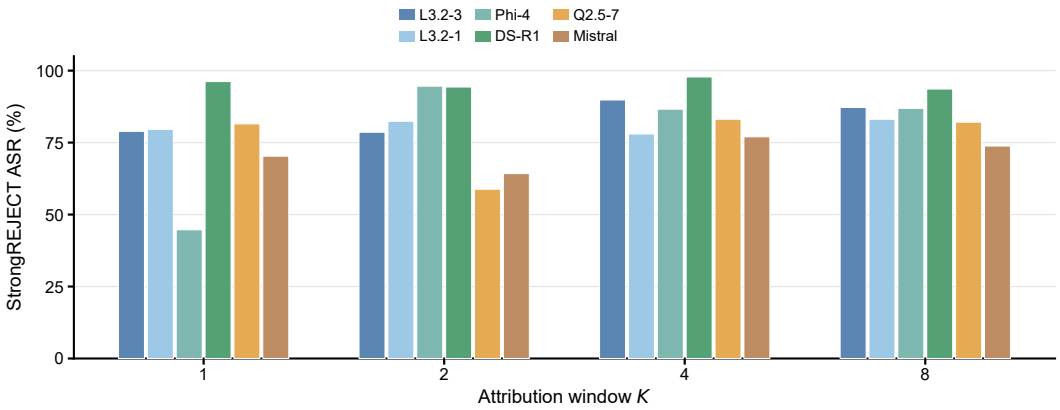


Figure 4: Sensitivity analysis of the prompt-boundary attribution window.

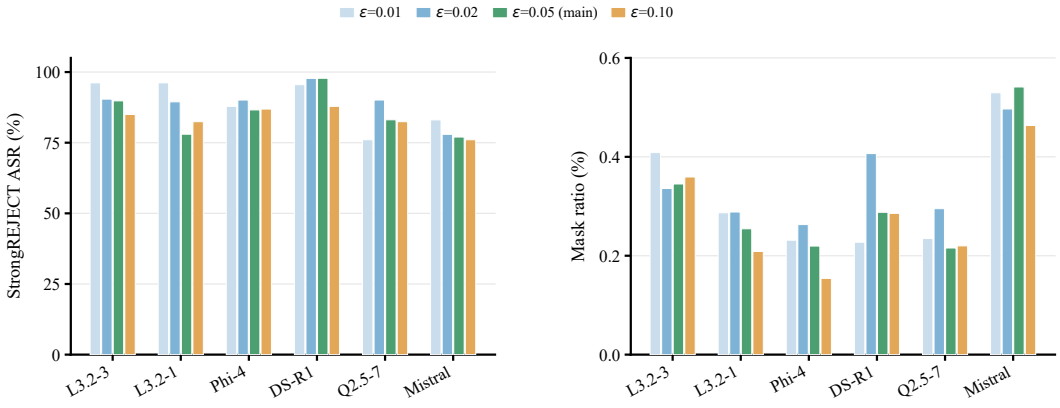


Figure 5: Sensitivity analysis of the behavioral-anchor budget.

Figure 4 shows that the effect of the attribution window is strongly model dependent and does not follow a simple monotonic pattern. Across the six representative models, the macro-average ASR for $K = 1$, $K = 2$, $K = 4$, and $K = 8$ is 75.2%, 78.8%, 85.4%, and 84.5%, respectively, with $K = 4$ achieving the highest average ASR. Llama-3.2-3B, DeepSeek-R1-Distill-Llama-8B, Qwen2.5-7B, and Mistral-7B-Instruct-v0.2 attain their highest ASR at $K = 4$, reaching 89.8%, 97.8%, 83.1%, and 77.0%, respectively. Phi-4-mini reaches 94.6% at $K = 2$, whereas Llama-3.2-1B reaches 83.1% at $K = 8$.

These results show that increasing the attribution window does not necessarily increase attack success rate. A window that is too small may not capture sufficient refusal-trajectory information, whereas a window that is too large may introduce contextual activations unrelated to refusal generation. The main experiments retain $K = 4$ as part of the shared configuration for consistent cross-model comparison.

A.2.2 Behavioral-Anchor Tolerance

The behavioral-anchor tolerance controls the allowed ASR decrease during coordinate-block deletion. A smaller tolerance imposes stricter behavioral constraints on deletion operations, whereas a larger tolerance permits greater ASR variation and may yield a sparser final mask. We vary $\epsilon \in \{0.01, 0.02, 0.05, 0.10\}$ while keeping all other main configurations unchanged.

As shown in Figure 5, stricter tolerance settings generally retain higher attack effectiveness, although final mask sizes do not vary strictly monotonically. Across the six models, the average ASR at $\epsilon = 0.01$, 0.02, 0.05, and 0.10 is 89.3%, 89.3%, 85.4%, and 83.4%, respectively. The corresponding average mask ratios are 0.32%, 0.35%, 0.31%, and 0.28%. These results reflect the behavioral–sparsity trade-off in conditional compression: as the tolerance is relaxed, the search can remove

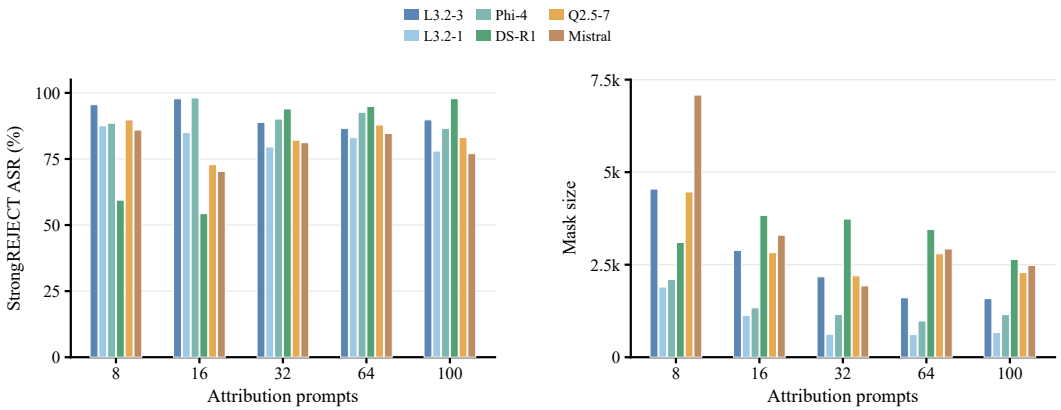


Figure 6: Sensitivity analysis of the attribution prompt budget.

more masked coordinates and typically obtains a smaller final static mask, while ASR may decrease accordingly.

Different models respond differently to the behavioral-anchor tolerance. For example, Llama-3.2-3B attains 96.2% ASR at $\epsilon = 0.01$, whereas the main configuration with $\epsilon = 0.05$ retains 89.8% ASR with a mask ratio of 0.35%. Phi-4-mini maintains ASR above 86.6% under all four settings, while its mask ratio decreases from 0.26% to 0.15%. Together, these results show model-specific ASR–sparsity trade-offs rather than a single universally optimal tolerance. To maintain a consistent cross-model protocol and avoid per-model tuning, the main experiments use $\epsilon = 0.05$ as a shared operating point.

A.2.3 Attribution Prompt Budget

The attribution prompt budget determines the number of attack prompts used to collect natural refusal trajectories during candidate localization. A smaller budget reduces offline construction cost, but stable voting may then cover only a limited set of refusal patterns. A larger budget provides more trajectory evidence and may alter both the candidate pool and the compressed final static mask. We construct candidate pools from 8, 16, 32, 64, and 100 JailbreakBench prompts while keeping stable voting, conditional compression, behavioral anchoring, and StrongREJECT evaluation unchanged.

Figure 6 shows that attack effectiveness does not change monotonically with the attribution prompt budget. The average ASR across the six models is 84.5%, 79.7%, 85.9%, 88.3%, and 85.4% for budgets of 8, 16, 32, 64, and 100 prompts, respectively. Although 64 prompts achieve the highest average ASR across these models, the budget yielding the highest ASR differs by model. Llama-3.2-3B and Phi-4-mini attain 97.8% and 98.1%, respectively, with 16 prompts. In contrast, DeepSeek-R1-Distill-Llama-8B attains 97.8% with 100 prompts. Its high baseline ASR on JBB-100 leaves fewer naturally generated refusals available for trajectory attribution, so a larger prompt budget yields more reference trajectories for stable voting,which may explain why a larger prompt budget is beneficial for this model.

In contrast, final mask size exhibits a clearer decreasing trend as the attribution prompt budget increases. The average mask size decreases from 3,866 coordinates with 8 prompts to 1,802 coordinates with 100 prompts, a reduction of 53.4%. This trend is particularly evident for Llama-3.2-3B, Llama-3.2-1B, and Mistral-7B-Instruct-v0.2, whose mask sizes decrease from 4,544, 1,896, and 7,084 coordinates to 1,583, 667, and 2,483 coordinates, respectively.

Because ASR and final mask size form a trade-off, we do not select the construction budget by ASR alone. Although 64 prompts yield a slightly higher average ASR than 100 prompts, the full JBB-100 configuration produces the smallest average mask and avoids selecting the main configuration from an arbitrary prompt subset. We therefore use JBB-100 for candidate-pool construction to maintain a consistent and reproducible offline construction protocol.