RESAT: TRAINING LLMS RESISTANT AGAINST REFUSAL ABLATION ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-weight language models allow downstream users to modify internal computations as well as prompts, creating a distinct challenge for preserving learned safeguards. We investigate whether safety training protects refusal and safety against the removal of multiple activation directions, rather than just one. We introduce *Refusal Subspace Ablation*, an attack that selects complementary directions and jointly removes their span throughout generation without updating model weights. Starting from one direction, it adds components outside the already selected span, choosing them by their effect under joint ablation. Our evaluations show that some current existing safety-trained models resist a selected single-direction attack yet remain vulnerable to joint ablation. To narrow this gap, we propose *Refusal Subspace Ablation Training*: mixed supervised fine-tuning followed by safety-training rounds that supervise safe responses under joint ablation and refresh candidate directions between rounds. We evaluate ReSAT on Llama-3-8B, Mistral-7B, Qwen2.5-14B, against RSA and other refusal-ablation attacks, with jailbreak attacks and utility test. In key settings, ReSAT achieves lower attack success rates than most other baselines while preserving clean MMLU accuracy. By connecting adaptive attack evaluation with safety training, our work offers a practical path toward more robust safeguards for open-weight LLMs.

1 INTRODUCTION

Open-weight language models enable local deployment, customization, and research, but also give downstream users access to the mechanisms responsible for refusal. Safeguarding these models therefore requires looking beyond adversarial prompts: users can fine-tune the weights or intervene directly in internal activations. Refusal-direction ablation makes this challenge concrete, suppressing refusal by removing an activation component without additional gradient-based training (Arditi et al., 2024; Wang et al., 2025). The circulation of modified models highlights the practical relevance of this threat: four Qwen3.8-27B derivatives advertised as *abliterated* or *uncensored* recorded a combined 571,199 Hugging Face downloads in the rolling month reported on 12 September 2026.[1]

Existing safety-training methods offer important defenses. MixAT combines continuous and discrete adversarial training (Dékány et al., 2025); ReFAT trains under dynamically refreshed refusal-feature ablation (Yu et al., 2025); and DeepRefusal uses probabilistic ablation of an offline refusal direction (Xie et al., 2025). Some resulting models retain high refusal rates under a selected single-direction attack. However, resistance to one direction does not establish resistance to the joint removal of complementary directions. Prior work, including Geometry (Wollschläger et al., 2025) and SOM (Piras et al., 2026), has explored multi-directional refusal ablation, although their computational cost is relatively high. We ask: *Do safety-trained models remain robust under adaptive joint ablation, and can training against cumulative interventions improve this robustness?*

We address the first question with *Refusal Subspace Ablation*. Keeping the attacked model fixed, it iteratively selects additional complementary directions according to their effect under joint ablation. Each selected direction remains part of the intervention, and the entire accumulated subspace is removed throughout generation. Applied to safety-trained checkpoints, this evaluation exposes weaknesses missed by a single-direction test. For example, HarmBench ASR on Llama DeepRefusal rises from 0.00% under single-direction ablation to 23.44% under our RSA (Table 1).

[1] Rolling monthly download counts from Hugging Face: Blackfrost-AI; orcarouter; huihui-ai; AEON-7.

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.

This failure mode motivates *Refusal Subspace Ablation Training*. After mixed supervised fine-tuning, we run SafeTrain rounds that supervise safe responses under joint removal of the accumulated refusal subspace. As training changes the model, we search for additional complementary directions between rounds, retaining the previously discovered span and expanding it when an eligible new direction is found. We evaluate the approach on Llama-3-8B, Mistral-7B, and Qwen2.5-14B. Our Llama H model achieves 11.20% HarmBench ASR under rank-four RSA; its unablated full-test MMLU is 66.49%, compared with 66.43% for the original instruction-tuned baseline. These results support training against cumulative interventions, while keeping robustness to existing jailbreak and refusal ablation attacks. Hence, our main contributions are:

- **Efficient multi-directional refusal ablation.** RSA discovers complementary refusal directions and jointly removes their span, achieving attack effectiveness competitive with or even higher than Geometry and SOM while requiring substantially less measured search time.
- **Robust defense against refusal ablation.** Between SafeTrain rounds, ReSAT refreshes candidate directions while retaining the previously selected subspace and learns safe responses under cumulative removal, outperforming evaluated baselines in key settings and transferring to Geometry and SOM.
- **Complementary to jailbreak defense.** ReSAT complements existing adversarial training methods to achieve robustness against refusal ablation and jailbreak attacks jointly, while preserving general utility.

2 Related Work

Refusal directions and causal activation interventions. Mechanistic studies of refusal identify low-dimensional activation features that separate harmful from harmless prompts and causally influence whether an instruction-tuned model refuses. A single residual-stream direction can be both necessary and sufficient for refusal across several model families: subtracting its component can induce harmful compliance, whereas adding it can elicit refusal on otherwise harmless inputs (Arditi et al., 2024). These findings establish activation intervention as a direct test of the internal mechanism, rather than merely another prompt-level jailbreak. They also motivate our central question: when safety training survives one known refusal direction, has it removed the vulnerability, or has the relevant computation migrated to a complementary subspace?

Multi-directional refusal suppression. Refusal need not be confined to one vector: Geometry (Wollschläger et al., 2025) studies concept cones, representational independence, and joint interventions, while SOM (Piras et al., 2026) develops multi-directional refusal suppression. Our contribution is an adaptive strong attack with low cost and more efficient search that selects complementary directions under joint ablation, which can be integrated into safety training.

Adversarial safety training. A complementary approach trains against input- or representation-space adversaries. Continuous Adversarial Training (CAT) replaces expensive discrete inner-loop search with continuous embedding perturbations and jointly rehearses benign utility data (Xhonneux et al., 2024). Latent Adversarial Training (LAT) injects targeted perturbations into intermediate representations to suppress persistent harmful behavior (Sheshadri et al., 2024). MixAT combines continuous and discrete adversarial examples to improve transfer across jailbreak families (Dékány et al., 2025). These methods expand the threat models encountered during training, but robustness to adversarial prompts or learned latent perturbations does not by itself establish robustness to causal removal of the feature that mediates refusal. We therefore evaluate them under interventions that repeatedly search and remove the currently exposed refusal component.

Training against refusal-feature removal. ReFAT directly simulates refusal-feature ablation during training and periodically refreshes the feature as the model changes (Yu et al., 2025). DeepRefusal augments harmful examples with adversarial response prefixes and trains under probabilistic activation ablation, encouraging refusal to recover after the response has already entered a harmful trajectory (Xie et al., 2025). Both approaches explicitly target refusal-feature removal and are consequently the closest baselines to our setting. The key distinction is the attack depth considered during optimization: a current or offline rank-one direction can become obsolete after the model reorganizes its computation. Between SafeTrain rounds, we extract fresh candidates from the updated checkpoint and search for a complementary direction while retaining the previously selected subspace.

3 REFUSAL SUBSPACE ABLATION

Resistance to a single refusal direction does not establish resistance to removing several directions together. A model may still refuse after one component is removed, yet lose that behavior when complementary components are removed jointly. We therefore construct an attack that progressively expands the ablated subspace while retaining previously selected directions. The key is to evaluate each new direction *together with the existing subspace*, rather than in isolation, while constraining changes on harmless prompts. We consider a white-box attacker who changes activations but keeps model weights fixed. The attack has two parts: a projection that removes a selected subspace, and an iterative search that decides which directions to include.

3.1 FROM ONE DIRECTION TO A JOINT SUBSPACE

Let $h \in \mathbb{R}^d$ be a hidden activation and let the columns of $B \in \mathbb{R}^{d\times k}$ be k orthonormal directions. The component $BB^\top h$ lies inside their span. Joint ablation subtracts this entire component, leaving only what lies outside:

$$A_B(h) = h - BB^\top h, \qquad B^\top B = I_k. \tag{1}$$

One column recovers rank-one ablation; an empty B gives the clean model. The same projection is applied at every decoder layer and token position, to the residual input and attention and MLP outputs, during both prompt processing and generation. Thus B describes the components removed together, not a list of independent single-direction tests.

3.2 FINDING COMPLEMENTARY DIRECTIONS BY JOINT EVALUATION

Step 1: extract candidates. Following prior work (Arditi et al., 2024), we compare mean activations on harmful and harmless prompts at candidate source layers and prompt positions. Their differences form a pool $\mathcal{C}^M$ for the fixed model M. These are candidate directions, not yet verified interventions. The source location specifies where a candidate is measured, not where it will be removed.

Step 2: remove already covered components. Starting from $B_0 = \varnothing$, suppose B_{k-1} contains the directions retained so far. Adding another vector within the same span would remove nothing new. We therefore subtract each candidate's projection onto that span, normalize its remaining component, and form a trial expanded basis:

$$\begin{aligned} r_k(c) &= (I_d - B_{k-1}B_{k-1}^\top)c, \\ v_k(c) &= r_k(c)/\|r_k(c)\|_2, \qquad Q_k(c) = \mathrm{Basis}([B_{k-1}, v_k(c)]). \end{aligned} \tag{2}$$

Here Basis returns an orthonormal basis; negligible residuals are excluded. Orthogonalization avoids redundant directions, but does not establish that the new intervention suppresses refusal.

Step 3: evaluate the full intervention and retain one expansion. We temporarily ablate each complete trial basis $Q_k(c)$ and score the model on harmful validation prompts. For a selected set R of refusal-probe tokens, S_R averages the log-odds of generating a token in R at the first assistant-output position; lower values favor refusal suppression. To limit nonspecific disruption, eligible candidates $\mathcal{E}_k$ must also satisfy a KL bound on the full first-token distribution for harmless prompts, relative to the clean model. We select

$$c_k \in \arg\min_{c\in\mathcal{E}_k} S_R(Q_k(c); M), \qquad B_k = Q_k(c_k). \tag{3}$$

The old directions remain active while the next expansion is tested. Raw clean candidates are cached for this fixed checkpoint; residuals and joint scores are recomputed at each step. Search stops at the requested rank or when no eligible candidate remains. The token score is only a selection proxy: generated refusal, utility, and semantic safety are evaluated separately. Appendix A specifies the exact score, probes, eligibility filters, and seed protocols.

4 RESAT: **RE**FUSAL **S**UBSPACE **A**BLATION **T**RAINING

We train the model to produce safe responses while accumulated refusal subspace is jointly removed. Our configuration uses mixed SFT initialization followed by iterative joint-ablation safety training (Figure 1). Only LoRA parameters θ are updated; pretrained weights remain frozen (Hu et al., 2021).

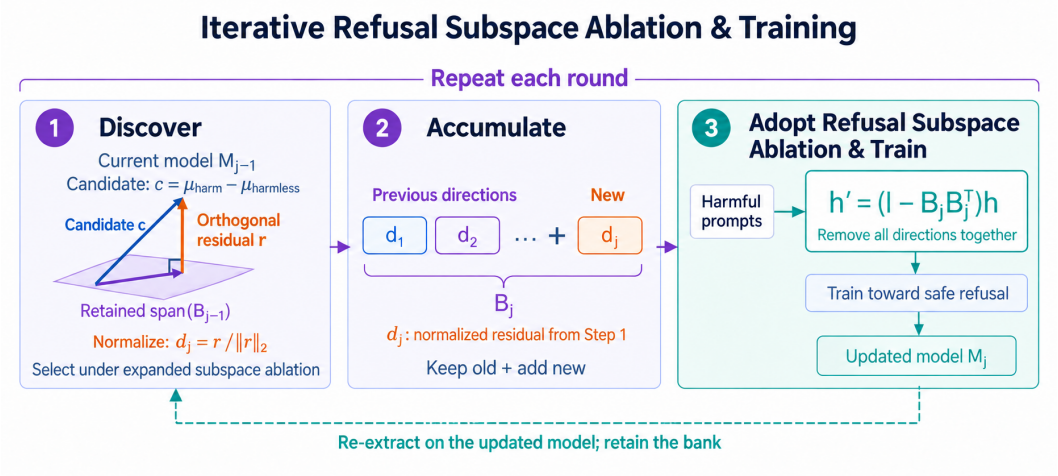


Figure 1: **ReSAT training loop.** RSA expands the retained direction bank, and SafeTrain learns safe responses under joint ablation. Candidate directions are refreshed between rounds.

4.1 Iterative joint-ablation training

Starting from the mixed-SFT adapter $\theta^{(0)}$ (Section 4.2), initialize $B_0 = \varnothing$. Before round j, freeze the current checkpoint $M_{\theta^{(j-1)}}$, extract fresh clean candidates, and apply the residualization and joint selection in Section 3 against the retained bank B_{j-1}. If an eligible new direction d_j is found, update

$$\begin{aligned} B_j &= \mathrm{Basis}([B_{j-1}, d_j]), \\ \theta^{(j)} &= \mathrm{Optimize}\big(\theta^{(j-1)}, \mathcal{L}_j(\,\cdot\,; B_j)\big). \end{aligned} \tag{4}$$

The complete B_j stays fixed within the round, while gradients pass through its projection to the LoRA parameters. If no eligible new direction is found, the original H procedure stops and retains the last completed round. Explicit fixed-bank variants instead continue with $B_j = B_{j-1}$; their rank does not increase.

A shared response loss. All training conditions supervise the assistant's target response. For target y, intervention B, and detached input perturbation $\bar{\delta}$, define

$$\ell_{\mathrm{resp}}(\theta; x, y; B, \bar{\delta}) = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log p_{\theta, B, \bar{\delta}}(y_t \mid x, y_{<t}). \tag{5}$$

Here x is the prompt, y_t the target token at position t, $y_{<t}$ its target prefix, and $|y|$ the supervised response length. $p_{\theta,B,\bar{\delta}}$ is the model's conditional distribution under ablation B and input perturbation $\bar{\delta}$; outer optimization treats $\bar{\delta}$ as constant. We average over response tokens, then examples, excluding user and padding tokens. $B = \varnothing$ and $\bar{\delta} = 0$ denote no ablation and no input attack, respectively. The training objective has five terms:

$$\begin{aligned} \mathcal{L}_j = \lambda_{\mathrm{clean}}\mathcal{L}_{\mathrm{clean}} + \lambda_{\mathrm{joint}}\mathcal{L}_{\mathrm{joint}} + \lambda_{\mathrm{cont}}\mathcal{L}_{\mathrm{cont}} \\ + \lambda_{\mathrm{pap}}\mathcal{L}_{\mathrm{pap+cont}} + \lambda_{\mathrm{help}}\mathcal{L}_{\mathrm{help}}. \end{aligned} \tag{6}$$

Each term is an average of Equation 5 over its presentation distribution; the nonnegative λ coefficients retain the original sampling shares and sum to one.

Clean refusal and joint-ablation refusal. For harmful prompts x with safe refusal targets y_{ref}, $\mathcal{L}_{\mathrm{clean}}$ averages $\ell_{\mathrm{resp}}(\theta; x, y_{\mathrm{ref}}; \varnothing, 0)$, maintaining refusal without intervention. $\mathcal{L}_{\mathrm{joint}}$ instead averages $\ell_{\mathrm{resp}}(\theta; x, y_{\mathrm{ref}}; B_j, 0)$, directly teaching refusal when the *entire accumulated subspace* is removed. The two terms use the same normalized harmful-prompt distribution and targets, but different forward computations.

Input attacks combined with joint ablation. $\mathcal{L}_{\text{cont}}$ averages $\ell_{\text{resp}}(\theta; x, y_{\text{ref}}; B_j, \bar{\delta}_x)$ after continuous perturbation search on user-token embeddings. $\mathcal{L}_{\text{pap+cont}}$ first samples a persuasive adversarial prompt (PAP) rewrite $\widetilde{x} = P(x; \xi)$, where P is the rewriting procedure and ξ its sampling randomness, then averages $\ell_{\text{resp}}(\theta; \widetilde{x}, y_{\text{ref}}; B_j, \bar{\delta}_{\widetilde{x}})$ over examples and rewrites. Both branches keep B_j active during inner attack search and outer training. The inner search increases the likelihood of an unsafe target prefix; the outer update always supervises the safe refusal. Perturbations are aligned to the outer input and detached: neither the attack optimizer nor direction selection is differentiated through.

Helpful responses on benign inputs. We group utility and hard-benign supervision under $\mathcal{L}_{\text{help}}$ while retaining their distinct roles. For each pool $a \in \{\text{util}, \text{hb}\}$, define

$$\mathcal{L}_a(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_a} \ell_{\text{resp}}(\theta; x, y; \varnothing, 0). \tag{7}$$

Here $\mathcal{D}_{\text{util}}$ contains general-task examples with helpful targets, so $\mathcal{L}_{\text{util}}$ preserves general capability. $\mathcal{D}_{\text{hb}}$ contains harmless requests prone to over-refusal, also paired with helpful rather than refusal targets; $\mathcal{L}_{\text{hb}}$ teaches the model to answer these requests normally. Each loss averages the response-normalized NLL within its own pool, without ablation or input perturbation. The grouped contribution is exactly

$$\begin{aligned} \lambda_{\text{help}}\mathcal{L}_{\text{help}} &= \lambda_{\text{util}}\mathcal{L}_{\text{util}} + \lambda_{\text{hb}}\mathcal{L}_{\text{hb}}, \\ \lambda_{\text{help}} &= \lambda_{\text{util}} + \lambda_{\text{hb}}. \end{aligned} \tag{8}$$

For SafeTrain, $\lambda_{\text{util}} = 0.35$ and $\lambda_{\text{hb}} = 0.20$ give $\lambda_{\text{help}} = 0.55$. This grouping preserves the six-condition recipe and its relative weights. There is no extra KL penalty in the outer loss: KL constrains direction selection only. The sampling shares, round count, branch definitions, token alignment, and optimization mechanics are in Appendix B.1.

4.2 Mixed SFT initialization

Mixed SFT aims to concentrate refusal around an identifiable direction while obscuring the base model's original direction, making attacks that reuse it less effective. Starting from the instruction-tuned checkpoint with a fresh LoRA adapter, we combine safe-refusal targets for harmful inputs with helpful answers for utility and hard-benign prompts. Harmful branches apply continuous attacks to both original and PAP-rewritten prompts (Zeng et al., 2024); no direction ablation is used during SFT. The resulting $\theta^{(0)}$ initializes SafeTrain, which continues the same adapter under cumulative direction removal. Appendix F gives the SFT composition; Section 5.5 reports the ablations.

5 Results

5.1 Experimental Setup

Models and evaluation sources. The three-backbone comparison uses Llama-3-8B-Instruct (Llama Team, AI @ Meta, 2024), Mistral-7B-Instruct (Jiang et al., 2023), Qwen2.5-14B-Instruct (Qwen et al., 2024) Evaluation cohorts are documented in Appendix C. We additionally rerun ReSAT with three independent seeds for each of the three model families; the results are consistent with the main-table findings, with full per-seed results reported in Appendix I.

Refusal-Subspace Ablation evaluation. After training, we freeze the checkpoint and construct evaluation subspaces on the safety-trained model. The first direction is selected from the safety-trained model checkpoint or transferred from the unmodified instruction-tuned base model. We then add complementary directions found on the evaluated checkpoint and jointly remove their accumulated span, expanding from one to at most four directions under the declared constraints. Search stops when no eligible direction remains under KL constraints.

Jailbreak attacks. NoAtk denotes the unmodified harmful-prompt control, not a separate attack method. We evaluate AutoDAN genetic prompt search (Liu et al., 2024), GCG (Zou et al., 2023), and PAIR (Chao et al., 2023). HumanJB uses the released HarmBench human-written jailbreak templates (Mazeika et al., 2024), sampled five per behavior in our evaluation. Inp. denotes a fixed transfer corpus of diffusion-inpainting attacks (Lüdke et al., 2025): we use the third-party Circuit Breakers attack-completions release (samuelsimko, 2026), not a new run of the original paper's complete search procedure. Soft uses the enhanced continuous soft-prompt implementation (Schwinn

Table 1: Refusal-ablation performance of base models and their safety-trained variants (%). Protocols and cohorts are specified in Appendix C. Refusal Rate evaluation is in Appendix E.

	ASR ↓					MMLU ↑				
Model	Clean	Single	Geometry	SOM	RSA (Ours)	Clean	Single	Geometry	SOM	RSA (Ours)
Llama-3-8B										
Base	0.78	92.19	98.74	79.25	92.19	**67.50**	64.75	66.00	**66.00**	64.25
LAT	**0.00**	85.16	94.97	80.50	86.72	66.50	64.75	63.75	64.25	62.75
CAT	**0.00**	27.34	49.69	80.50	45.31	65.75	**65.75**	61.00	64.75	64.00
MixAT	**0.00**	78.91	83.02	87.42	82.03	62.25	63.50	57.25	62.75	60.25
ReFAT	**0.00**	89.84	97.48	72.96	92.19	66.25	**65.75**	64.25	64.75	**65.25**
DeepRefusal	**0.00**	**0.00**	**0.00**	22.64	23.44	65.50	64.75	**66.75**	63.00	64.25
Ours(ReSAT)	**0.00**	4.00	**0.00**	**0.00**	**11.20**	65.75	63.75	65.50	**66.00**	64.75
Mistral-7B										
Base	29.69	69.53	29.69	88.68	69.53	**60.50**	**59.50**	60.50	52.00	**59.50**
LAT	3.91	62.50	18.24	81.13	81.25	54.75	55.25	45.00	32.00	54.00
CAT	1.56	10.94	76.73	63.52	75.78	54.50	52.50	44.75	29.75	53.25
MixAT	0.78	6.25	**0.00**	**4.40**	65.62	54.50	54.75	47.50	47.25	55.75
ReFAT	**0.00**	28.91	87.42	49.69	28.91	59.00	57.25	53.00	50.50	57.25
DeepRefusal	3.12	78.91	50.94	89.94	78.91	59.00	57.75	**55.75**	**52.75**	57.75
Ours(ReSAT)	**0.00**	**1.60**	**0.00**	28.57	**20.00**	57.25	57.25	55.50	47.25	57.25
Qwen2.5-14B										
Base	1.56	79.69	96.23	98.11	79.69	76.50	76.50	**78.50**	76.75	76.50
LAT	**0.00**	47.66	83.65	95.60	47.66	76.75	76.00	76.25	76.00	76.00
CAT	0.78	1.56	58.49	3.14	**1.56**	74.75	75.50	75.50	73.00	75.50
MixAT	**0.00**	**0.00**	**1.26**	**0.00**	29.69	76.50	76.50	77.25	**78.00**	70.25
ReFAT	**0.00**	27.34	98.11	96.86	64.06	76.75	76.50	**78.50**	76.50	75.50
DeepRefusal	**0.00**	61.72	96.23	86.16	72.66	76.25	76.25	75.75	76.50	75.75
Ours(ReSAT)	**0.00**	1.50	7.14	22.73	8.80	78.50	**79.00**	**78.50**	76.75	**79.00**

& Geisler, 2024). RFA uses the layerwise, mean-centered refusal-feature ablation operator evaluated by ReFAT (Yu et al., 2025); it is distinct from our adaptive joint-subspace attack.

Computational cost. Single-run wall-clock times on NVIDIA GH200 120GB GPUs are 154.2 minutes for SFT, 387.6 for SafeTrain, and 619.2 end to end for Llama-3-8B ReSAT. Training uses two GPUs, the second for PAP generation. Figure 2 compares seven Llama checkpoints; RSA has the lowest measured core-search time, and on Base RSA/Geometry/SOM take 4/101/1,310 minutes. Its low cost enables repeated SafeTrain direction refreshes, whereas the other joint-ablation searches are substantially costlier. Full accounting and attack-specific timing boundaries are in Appendix D.

Scoring and utility. Attack Successful Rate (ASR) is measureed by calculating average of 3 classifier: HarmBench (Mazeika et al., 2024), Llama Guard 2 (Llama Team, 2024), and StrongREJECT (Souly et al., 2024), reported as percentages. MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) measure capability; XSTest (Röttger et al., 2024) measures benign compliance.

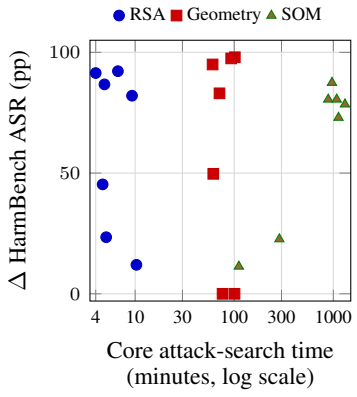


Figure 2: Core search time versus HarmBench ASR increase on seven Llama checkpoints; details are in Appendix D.2.

Clean has no direction intervention; lexical RR is Refusal Rate reported as another supplementary metrics in Appendix E. Best values are **bold**; second-best are underlined.

5.2 Resistance to Refusal Ablation Attacks

ReSAT provides strong defense against a range of refusal-ablation attacks. Table 1 shows strong resistance to both single-direction and different types of cumulative joint ablation across 3 model backbones, with substantially lower ASR than several existing safety-training baselines.

Although ReSAT uses RSA to optimize SafeTrain, it also provides strong defense against Geometry and SOM. For example, Llama Ours achieves 0.00% ASR under both attacks (Table 1), showing that robustness learned with RSA can be transferred to other attacks. This transfer demonstrates good generalization and robustness across different models

5.3 Attack Performance

RSA, our joint-ablation attack, is highly effective against existing safety-trained models. Models that resist single-direction ablation can remain vulnerable to joint removal of multiple directions. For example, Mistral MixAT's ASR rises from 6.25% under Single to 65.62% under rank-three RSA; similar weaknesses appear in CAT, DeepRefusal and other safety-training baselines (Table 1). Relaxing the KL constraint can strengthen RSA but may incur greater utility loss, with effects varying across models (Appendix H).

RSA is more time-efficient than Geometry and SOM in our measured runs (Figure 2). In the Llama timing study, our core search takes 10.3 minutes, versus 76.30 for Geometry and 111.33 for SOM. Although search budgets differ, RSA's low measured cost makes the attack practical and enables repeated direction refreshes during SafeTrain. Together, these results show that single-direction robustness is not a reliable proxy for subspace robustness, while RSA exposes this gap at substantially lower measured search cost.

5.4 Utility Preservation and Jailbreak Robustness

ReSAT largely preserves general capability after SafeTrain. In Table 2, MMLU for Llama Ours (66.49%) and Base (66.43%) ; for Mistral Ours(59.22%)/Base(58.40%); Qwen ReSAT(ours) achieves 78.93% versus base 78.64% with also GSM8K improves from 78.85% to 85.90% relativr to base. ReSAT exhibits over-refusal on XSTest, a limitation also observed in other baseline CAT, MixAT, and DeepRefusal. However, On Llama-3-8B, Our ReSAT achieves higher benign compliance.

ReSAT also defends effectively against several jailbreak attacks, outperforming ReFAT and DeepRefusal in multiple tested settings. On Llama, Ours's GCG/PAIR/Inpainting mixed safety scores (lower is better) are 1.05/1.19/47.31, compared with 28.72/28.29/83.00 for ReFAT and 10.65/6.01/52.26 for DeepRefusal. On Mistral, Ours's AutoDAN score is 4.12, versus 54.00 for ReFAT and 68.82 for DeepRefusal (Table 2). These gains vary across attacks, but they show promising robustness for our safe-training method ReSAT.

5.5 Ablation Study

SFT initialization affects the robustness compliance trade-off. Table 3(i) compares original, mixed, and no SFT initialization under the same three-round SafeTrain mixture; total training compute differs. Mixed SFT improves benign compliance but increases observed RSA ASR, while omitting SFT largely removes RSA resistance (Appendix F).

Input-attack training improves measured jailbreak robustness. Group (ii) removes continuous/PAP attacks from both SFT and SafeTrain while preserving hard-benign exposure and update budgets. Removing these attacks raises the Inpainting score from 47.31% to 67.77%, testing their combined contribution across both stages.

Hard-benign supervision improves benign compliance. Group (iii) replaces hard-benign examples with ordinary utility while holding the harmful branches and update budget fixed. Adding hard-benign supervision raises XSTest compliance from 9% to 53%, at the cost of higher rank-four RSA ASR.

Table 2: Jailbreak robustness and utility of base models and their safety-trained variants (%).

	Jailbreak safety score ↓								Utility		
Model	NoAtk	Auto DAN	Human JB	GCG	PAIR	Inp.	Soft	RFA	MMLU ↑	GSM 8K↑	XS Test↑
Llama-3-8B											
Base	1.57	18.20	2.20	22.97	28.70	86.61	79.20	62.31	66.43	**76.19**	**99.00**
LAT	0.04	1.74	0.03	7.32	23.30	62.92	**75.55**	70.35	**66.55**	66.26	67.00
CAT	**0.02**	**0.02**	**0.02**	6.57	32.51	59.07	80.26	**33.12**	65.08	70.28	26.00
MixAT	**0.02**	**0.02**	**0.02**	13.82	1.95	66.79	77.34	60.97	63.70	71.80	49.00
ReFAT	0.44	21.43	1.16	28.72	28.29	83.00	82.67	49.06	66.54	74.07	91.00
DeepRefusal	0.03	**0.02**	**0.02**	10.65	6.01	52.26	77.26	57.75	65.99	71.87	48.00
Ours(ReSAT)	0.04	0.04	0.04	**1.05**	**1.19**	**47.31**	82.12	57.25	66.49	73.09	53.00
Mistral-7B											
Base	37.79	79.02	65.56	79.96	62.60	94.51	83.28	78.31	58.40	41.85	95.00
LAT	0.60	47.53	10.51	28.82	61.65	91.18	81.51	79.69	52.79	0.00	67.00
CAT	0.40	2.32	0.33	20.99	47.33	86.27	**80.43**	76.94	51.53	33.97	86.00
MixAT	0.05	**0.02**	**0.02**	**12.50**	**13.14**	77.46	80.78	72.35	52.93	34.27	62.00
ReFAT	0.07	54.00	6.53	67.40	38.67	87.76	82.77	75.77	58.09	27.29	72.00
DeepRefusal	1.77	68.82	14.88	69.39	59.67	90.88	82.93	78.24	58.13	41.32	**97.00**
Ours(ReSAT)	**0.04**	4.12	0.07	53.83	30.49	**64.63**	80.51	**47.29**	**59.22**	**42.99**	57.00
Qwen2.5-14B											
Base	3.04	29.72	11.29	12.89	31.01	87.54	84.46	29.28	78.64	78.85	**99.00**
LAT	0.08	0.04	0.06	**0.26**	21.50	70.24	74.20	**25.01**	78.69	56.25	69.00
CAT	0.16	**0.02**	**0.02**	1.62	35.81	64.68	62.25	39.89	77.45	82.94	44.00
MixAT	**0.02**	**0.02**	**0.02**	3.36	**5.50**	**56.64**	**54.60**	40.33	78.23	83.09	41.00
ReFAT	0.16	2.08	1.60	3.45	27.34	83.78	85.29	31.67	**78.95**	79.91	80.00
DeepRefusal	0.12	0.03	0.08	32.58	43.18	77.74	84.81	27.83	78.35	82.79	90.00
Ours(ReSAT)	0.04	4.58	0.65	52.48	44.21	80.60	83.75	52.07	78.93	85.90	47.00

RSA resistance improves mainly between the first two rounds. Group (iv) compares checkpoints after rounds one, two, and three of the sametraining run. Rank-four ASR falls from 35.20% to 12.00% and 11.20%, with only a one-response difference between the final two rounds (Appendix F).

Cumulative-direction training improves resistance over repeated single-direction training. Group (v) compares repeated ablation of the first fixed direction d_1 with cumulative direction ablation during training, using the same final-model evaluation protocol. Cumulative training lowers rank-four ASR from 31.20% to 11.20%.

Training-time RSA improves refusal-ablation resistance in a matched control. Group (vi) disables only training-time direction ablation while retaining the initialization, sample quotas, continuous/PAP attacks, and update budget; both models are evaluated with RSA. At the shared attack rank three, training-time RSA lowers ASR from 61.60% to 12.00%, with lower benign compliance.

5.6 Mechanism Analysis

SafeTrain progressively suppresses attacks from known direction banks. Appendix Figure 4A shows that the causal attack effect of each identified bank decreases across successive checkpoints. For example, the ASR increase induced by B_1 falls from 51.2 pp on M_0 to 8.0, 1.9, and 2.4 pp on M_1, M_2, and M_3, demonstrating robustness acquisition against known interventions.

Retaining direction history improves final robustness. We compare refresh+accumulate with a matched refresh+replace schedule that re-estimates a direction every round but discards the preceding bank. Replacement reaches 29.07% HarmBench ASR on M_3, compared with 11.20% under accumulation (Figure 3), a paired difference of 17.87 pp (95% CI: $[14.13, 21.87]$).

Table 3: Llama-3-8B ablations (%). RSA reports HarmBench ASR. Setup and mixture details are in Appendix F.

Setup	MMLU↑	XSTest↑	AutoDAN↓	GCG↓	Inpainting↓	RSA↓
(i) SFT initialization: same SafeTrain mixture						
Original SFT	65.88	9.00	1.76	2.78	49.93	**7.20**
Mixed SFT	66.42	46.00	**0.04**	**2.35**	**46.38**	12.00
No initial SFT	**66.56**	**97.00**	12.71	18.91	84.41	93.60
(ii) Input attacks: matched hard-benign exposure and updates						
No input attacks	**66.53**	**65.00**	9.94	7.86	67.77	16.80
With input attacks	66.49	53.00	**0.04**	**1.05**	**47.31**	**11.20**
(iii) Hard-benign supervision: ordinary-utility control						
No Hard Benign	65.61	9.00	**0.04**	5.81	**45.63**	**4.00**
With Hard Benign	**66.49**	**53.00**	**0.04**	**1.05**	47.31	11.20
(iv) Training rounds: successive checkpoints of the same run						
Ours after round 1	66.38	**56.00**	**0.04**	3.50	42.63	35.20
Ours after round 2	**66.54**	45.00	**0.04**	**0.04**	**40.34**	12.00
Ours after round 3	66.49	53.00	**0.04**	1.05	47.31	**11.20**
(v) Complementary directions: fixed versus cumulative ablation						
Repeated d_1	66.46	**60.00**	**0.04**	5.01	48.17	31.20
Cumulative directions (Ours)	**66.49**	53.00	**0.04**	**1.05**	**47.31**	**11.20**
(vi) Training-time RSA: matched initialization, quotas, and updates						
Without training RSA	**66.49**	**79.00**	**0.04**	3.37	53.60	61.60
With training RSA	**66.49**	53.00	**0.04**	**1.05**	**47.31**	**11.20**

Cumulative history prevents rediscovery of covered directions. After projection against the historical span, accumulated d_2 and d_3 retain 76.0% and 99.9% of their squared energy, whereas replacement candidates retain only 0.185% and 0.043% (Appendix Figure 4C). Thus, accumulation exposes complementary information while replacement largely rediscovers covered directions.

The complementary directions are causally active. Adding accumulated d_2 and d_3 to their preceding banks increases held-out ASR by 26.93 and 4.27 pp, respectively, while replacement candidates provide no positive marginal gain (Appendix Figure 4D). The newly exposed information therefore causally controls refusal rather than merely reflecting representation rotation.

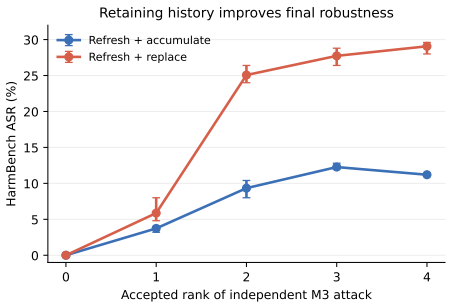


Figure 3: History retention lowers HarmBench ASR across accepted ranks of an independent M_3 attack compared with replacing the previous direction bank (%; lower is better): History retention improves robustness

6 Conclusion

Resistance to a single refusal direction does not ensure safety under joint ablation. RSA adaptively discovers and jointly removes complementary directions, achieving competitive attack effectiveness with substantially lower measured search cost than Geometry and SOM. ReSAT combines mixed SFT with safety training under cumulative ablation. Across model families, ReSAT lowers attack success rates relative to evaluated baselines in key settings while preserving general capability. Its robustness extends to other refusal-ablation and several jailbreak attacks, complementing existing adversarial training for open-weight models. Our mechanism analysis further shows that retaining the cumulative direction history prevents rediscovery of already-covered features and exposes causally active complementary directions.

AI USE STATEMENT

We used generative AI tools to assist with translation, language editing, drafting and revising portions of the manuscript, interpreting experimental results, and checking consistency between the text and tables. These tools also assisted with LaTeX editing and formatting.

ETHICS STATEMENT

This work studies refusal-ablation attacks to evaluate and improve the safety of open-weight language models. RSA has dual-use implications: interventions that reveal vulnerabilities could also be misused to bypass safeguards and elicit harmful content. Our intended use is controlled safety evaluation and the development of defenses such as ReSAT. The reported robustness gains are limited to the evaluated models, attacks, and benchmarks and do not constitute a general safety guarantee. We encourage responsible use of these methods for authorized safety research and defensive evaluation.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083. Curran Associates, Inc., 2024. doi: 10.52202/079017-4322.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Csaba Dékány, Stefan Balauca, Robin Staab, Dimitar I. Dimitrov, and Martin Vechev. MixAT: Combining continuous and discrete adversarial training for LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96436–96469, 2025. doi: 10.52202/085713-2902.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f83cb637e159e789f5576ff6848874de-Abstract-Conference.html.

Llama Team. Meta Llama Guard 2. Model card, 2024. URL https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md.

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

David Lüdke, Tom Wollschläger, Paul Ungermann, Stephan Günnemann, and Leo Schwinn. Diffusion LLMs are natural adversaries for any LLM. *arXiv preprint arXiv:2511.00203*, 2025. URL https://arxiv.org/abs/2511.00203.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Giorgio Piras, Raffaele Mura, Fabio Brau, Luca Oneto, Fabio Roli, and Battista Biggio. SOM directions are better than one: Multi-directional refusal suppression in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32728–32736, 2026. doi: 10.1609/aaai.v40i39.40551. URL https://ojs.aaai.org/index.php/AAAI/article/view/40551.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

samuelsimko. RR (Circuit Breakers) attack completions with three-judge scores. Hugging Face dataset repository, 2026. URL https://huggingface.co/datasets/samuelsimko/rr-circuit-breakers-attack-completions. Experiment-pinned revision b863e72cd710a2eb55311b06b08f913e71791db6; dataset card accessed 14 September 2026.

Leo Schwinn and Simon Geisler. Revisiting the robust alignment of circuit breakers. *arXiv preprint arXiv:2407.15902*, 2024. URL https://arxiv.org/abs/2407.15902.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. *arXiv preprint arXiv:2407.15549*, 2024. URL https://arxiv.org/abs/2407.15549.

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, 2024. doi: 10.52202/079017-3984. URL https://papers.nips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *Advances in Neural Information Processing Systems*, volume 38, pp. 36471–36514, 2025. doi: 10.52202/085713-1086.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in LLMs with continuous attacks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1502–1530, 2024. doi: 10.52202/079017-0048.

URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0302fb83c62991efbccf0a003e4f5a92-Abstract-Conference.html.

Yuanbo Xie, Yingjie Zhang, Tianyun Liu, Duohe Ma, and Tingwen Liu. Beyond surface alignment: Rebuilding LLMs safety mechanism via probabilistically ablating refusal direction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17620–17632, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.956. URL https://aclanthology.org/2025.findings-emnlp.956/.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust LLM safeguarding via refusal feature adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s5orchdb33.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.773. URL https://aclanthology.org/2024.acl-long.773/.

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

A Detailed Refusal Ablation Protocol

A.1 Background: Naive Refusal Ablation

Let $h^M_{p,\ell}(x) \in \mathbb{R}^d$ denote the clean residual activation at the input of decoder layer ℓ for prompt x in model M. Here d is the hidden dimension, $p \in \mathcal{P}$ is a candidate prompt position relative to the assistant boundary, and $\ell \in \{0, \ldots, L-1\}$ is a zero-based layer index in a model with L decoder layers. Given harmful and harmless extraction sets $\mathcal{D}^{\text{ext}}_{\text{harm}}$ and $\mathcal{D}^{\text{ext}}_{\text{safe}}$, we first compute clean difference-in-means candidates:

$$c^M_{p,\ell} = \frac{1}{|\mathcal{D}^{\text{ext}}_{\text{harm}}|} \sum_{x \in \mathcal{D}^{\text{ext}}_{\text{harm}}} h^M_{p,\ell}(x) - \frac{1}{|\mathcal{D}^{\text{ext}}_{\text{safe}}|} \sum_{x \in \mathcal{D}^{\text{ext}}_{\text{safe}}} h^M_{p,\ell}(x). \tag{9}$$

For a nonzero candidate, let $d_{p,\ell} = c^M_{p,\ell}/\|c^M_{p,\ell}\|_2$. For any unit direction d, rank-one ablation removes its component from an intervened activation h:

$$A_d(h) = h - (h^\top d)d. \tag{10}$$

The source pair (p, ℓ) specifies where the candidate is extracted, not where it is removed. Unless stated otherwise, the same projection is applied at all sequence positions and decoder layers, at the residual input and the attention and MLP outputs, throughout prompt processing and generation. A rank-one attack keeps this one direction fixed; it is not a one-token or one-forward-pass intervention.

First-output-token selection. Candidate extraction and candidate selection are separate operations. Our naive selector fixes

$$R = \{t_I\}, \tag{11}$$

where t_I is the token ID corresponding to `I` at the first assistant-output position under the model's tokenizer and chat template. The probe is evaluated immediately after the native assistant boundary, without inserting `I` into the response. In particular, a refusal direction is not the embedding of `I`.

For a general set R of distinct probe token IDs, let $p_{M,Q}(\cdot \mid x)$ denote the distribution of the first token to be generated when the subspace with orthonormal basis Q is ablated; $Q = \varnothing$ denotes the clean model. Using harmful and harmless validation sets $\mathcal{D}^{\text{val}}_{\text{harm}}$ and $\mathcal{D}^{\text{val}}_{\text{safe}}$, we define

$$m_R(x; M, Q) = \log \frac{\sum_{t \in R} p_{M,Q}(y_1 = t \mid x)}{1 - \sum_{t \in R} p_{M,Q}(y_1 = t \mid x)}, \tag{12}$$

$$S_R(Q; M) = \frac{1}{|\mathcal{D}^{\text{val}}_{\text{harm}}|} \sum_{x \in \mathcal{D}^{\text{val}}_{\text{harm}}} m_R(x; M, Q), \tag{13}$$

$$K(Q; M) = \frac{1}{|\mathcal{D}^{\text{val}}_{\text{safe}}|} \sum_{x \in \mathcal{D}^{\text{val}}_{\text{safe}}} D_{\text{KL}}(p_{M,\varnothing}(\cdot \mid x) \,\|\, p_{M,Q}(\cdot \mid x)). \tag{14}$$

Thus m_R takes the log-odds of the *sum* of probe-token probabilities, and S_R averages these per-prompt log-odds. The KL is over the full vocabulary at the first output step, from the clean distribution to the distribution under the complete intervention, not between consecutive attack rounds. S_R is a direction-selection proxy, not the refusal rate (RR) measured from generated answers.

Let $\mathcal{E}_1(R; b_{\text{add}})$ be the eligible source pairs defined below for an empty initial basis, where $b_{\text{add}} \in \{0, 1\}$ records whether the optional activation-addition check is enabled. The naive attack selects

$$(p^\star, \ell^\star) \in \arg\min_{(p,\ell) \in \mathcal{E}_1(\{t_I\}; b_{\text{add}})} S_{\{t_I\}}([d_{p,\ell}]; M), \tag{15}$$

and retains that fixed rank-one projection throughout generation. The standalone `I`-selection implementation uses the KL and source-layer constraints without the addition check ($b_{\text{add}} = 0$). Addition-constrained runs are specified separately. The `Sorry` and `I+Sorry` selectors replace R by $\{t_{\text{Sorry}}\}$ and $\{t_I, t_{\text{Sorry}}\}$, respectively; they are probe variants, not different definitions of candidate extraction.

A.2 Refusal Subspace Ablation

Refusal Subspace Ablation adaptively selects complementary directions and jointly removes their span while keeping model weights fixed. Let $B_0 \in \mathbb{R}^{d\times 0}$ be empty. At step k, B_{k-1} has orthonormal columns spanning the components already removed. The clean candidate pool is

$$\mathcal{C}^M = \{c^M_{p,\ell} : p \in \mathcal{P},\ \ell = 0, \dots, L-1\}. \tag{16}$$

It is computed once from Equation 9 and can be cached for this checkpoint. We do not recompute harmful/harmless activation means under the current joint intervention. Instead, every step re-residualizes and jointly re-scores the clean candidates against the current span:

$$r_{k,p,\ell} = (I_d - B_{k-1}B_{k-1}^\top)c^M_{p,\ell}, \qquad \nu_{k,p,\ell} = \frac{r_{k,p,\ell}}{\|r_{k,p,\ell}\|_2}. \tag{17}$$

Here I_d is the $d \times d$ identity matrix; $r_{k,p,\ell}$ is the unnormalized residual, and $\nu_{k,p,\ell}$ is its unit vector, defined only for nonzero residuals. For each surviving candidate, form

$$Q_{k,p,\ell} = \mathrm{Basis}([B_{k-1}, \nu_{k,p,\ell}]), \tag{18}$$

where Basis returns an orthonormal basis for the columns' span. Both $S_R(Q_{k,p,\ell}; M)$ and $K(Q_{k,p,\ell}; M)$ evaluate the complete expanded intervention, not the new vector in isolation.

Optional activation-addition check. Some selection protocols additionally require a candidate to induce the refusal probe on harmless validation prompts. Let $p^+_{M;r,\ell}(\cdot \mid x)$ be the first-output-token distribution obtained by adding the *unnormalized* vector r to the residual input of source layer ℓ at all prompt positions. This is a separate clean-model forward pass with only the addition hook: the current subspace is *not* simultaneously ablated. Define

$$J_R(r, \ell; M) = \frac{1}{|\mathcal{D}^{\mathrm{val}}_{\mathrm{safe}}|} \sum_{x \in \mathcal{D}^{\mathrm{val}}_{\mathrm{safe}}} \log \frac{\sum_{t\in R} p^+_{M;r,\ell}(y_1 = t \mid x)}{1 - \sum_{t\in R} p^+_{M;r,\ell}(y_1 = t \mid x)}. \tag{19}$$

When enabled, the check uses $r = r_{k,p,\ell}$, not $\nu_{k,p,\ell}$, and requires $J_R \geq \eta$. It is a probe-based steering check rather than a generated-answer RR test.

Eligibility and joint selection. For nonzero raw candidates, let $\mathcal{E}_k(R; b_{\mathrm{add}})$ contain the source pairs satisfying all of

$$p \in \mathcal{P}, \qquad 0 \leq \ell < \lfloor (1 - \rho_{\mathrm{prune}})L \rfloor, \tag{20}$$

$$\|r_{k,p,\ell}\|_2 > 0, \qquad \frac{\|r_{k,p,\ell}\|_2}{\|c^M_{p,\ell}\|_2} \geq \epsilon, \qquad K(Q_{k,p,\ell}; M) \leq \tau, \tag{21}$$

$$b_{\mathrm{add}} = 0 \quad \text{or} \quad J_R(r_{k,p,\ell}, \ell; M) \geq \eta. \tag{22}$$

Here ρ_{prune} is the fraction of final source layers excluded, ϵ is the minimum retained residual fraction, and τ is the harmless first-token KL budget. We use $\rho_{\mathrm{prune}} = 0.2$ unless stated otherwise: for $L = 32$, the implementation admits $\ell < 25$, not $\ell < 25.6$. These are source-layer restrictions, not restrictions on where ablation is applied. The switch b_{add} and threshold η specify the optional check. At $k = 1$, $r_{1,p,\ell} = c^M_{p,\ell}$ and $Q_{1,p,\ell} = [d_{p,\ell}]$, which defines the naive eligibility set above (with $\epsilon \leq 1$).

The selected component and updated basis are

$$\begin{aligned} (p_k, \ell_k) &\in \arg\min_{(p,\ell)\in\mathcal{E}_k(R;b_{\mathrm{add}})} S_R(Q_{k,p,\ell}; M), \\ B_k &= \mathrm{Basis}([B_{k-1}, \nu_{k,p_k,\ell_k}]). \end{aligned} \tag{23}$$

Writing $d_k = \nu_{k,p_k,\ell_k}$, joint ablation then removes the entire accumulated span:

$$A_{B_k}(h) = h - B_k B_k^\top h. \tag{24}$$

The procedure repeats until the requested rank is reached or no candidate is eligible. It imposes no requirement that the proxy score or generated RR improve over the previous step, so neither is guaranteed to decrease monotonically. When the selector, prompt preprocessing and splits, candidate grid, and eligibility constraints are identical, the first step from $B_0 = \varnothing$ equals the corresponding naive attack; subsequent steps add orthogonal components and retain joint removal of all selected directions. The direction bank is only an implementation of this subspace.

Protocol scope. The general adaptive-matrix runner defaults to $b_{\text{add}} = 1$ and $\eta = 0$. Some Mistral continuation runs instead set $\eta = -10^9$, effectively disabling the addition filter; the standalone naive extractor omits it. Consequently, the label *strict* alone does not establish that two runs use identical filters: comparisons must retain their recorded selector, KL budget, and addition-filter setting. Training-time selection for Llama H uses the `I` probe, KL ≤ 0.2, and no addition check ($b_{\text{add}} = 0$); the default evaluation bound is KL ≤ 0.1. The historical F training selector also omits the addition check. Appendix B.1 gives the search settings. Section 5 evaluates refusal and utility under the selected interventions.

A.3 RSA Pseudocode

Algorithm 1 summarizes the executable RSA loop specified in Appendix A.

Algorithm 1 Refusal Subspace Ablation (RSA).

Require: Frozen model M; extraction sets $\mathcal{D}^{\text{ext}}_{\text{harm}}, \mathcal{D}^{\text{ext}}_{\text{safe}}$; validation sets $\mathcal{D}^{\text{val}}_{\text{harm}}, \mathcal{D}^{\text{val}}_{\text{safe}}$; maximum rank K; probe-token set R; KL budget τ
Ensure: Orthonormal refusal-subspace basis B

Extract the clean candidate pool $\mathcal{C}^M = \{c^M_{p,\ell}\}_{p,\ell}$ by difference in means
$B \leftarrow \varnothing$
for $k = 1, \ldots, K$ **do**
  $\mathcal{E} \leftarrow \varnothing$
  for $c \in \mathcal{C}^M$ **do**
    $r \leftarrow (I - BB^\top)c$
    if $\|r\|_2$ is negligible **then**
      continue
    end if
    $Q \leftarrow \text{Basis}([B, r/\|r\|_2])$
    if $K(Q; M) > \tau$ **then**
      continue
    end if
    Add $(S_R(Q; M), Q)$ to $\mathcal{E}$ ▷ apply the optional addition filter if enabled
  end for
  if $\mathcal{E} = \varnothing$ **then**
    break
  end if
  $B \leftarrow Q^\star$, where $(\cdot, Q^\star) \in \arg\min_{(s,Q)\in\mathcal{E}} s$
end for
return B; throughout generation jointly ablate $h \leftarrow h - BB^\top h$

B Training Details and Diagnostics

This appendix describes the core training protocol and two retained mechanistic diagnostics. Ablation setups and sampling shares are in Appendix F; shared evaluation definitions and cohorts are in Appendix C.

B.1 CORE SFT AND JOINT-ABLATION TRAINING PROTOCOL

Algorithm 2 ReSAT-H training.

Require: Base model M; mixed-SFT data $\mathcal{D}_{\mathrm{SFT}}$; SafeTrain data $\mathcal{D}$; rounds $J = 3$
Ensure: Last completed LoRA adapter and retained direction bank

$\theta^{(0)} \leftarrow \mathrm{MixedSFT}(M, \mathcal{D}_{\mathrm{SFT}})$ with all projections disabled
$B_0 \leftarrow \varnothing$
for $j = 1, \ldots, J$ **do**
  Freeze $M_{\theta^{(j-1)}}$ and refresh its clean RSA candidate pool
  $d_j \leftarrow \mathrm{RSASelectOne}(M_{\theta^{(j-1)}}, B_{j-1})$
  if no eligible d_j exists **then**
    break
  end if
  $B_j \leftarrow \mathrm{Basis}([B_{j-1}, d_j])$
  Set H/R/CR/PR/U/HB sampling shares to $10/22.5/6.25/6.25/35/20\%$
  Apply B_j in the R/CR/PR branches; keep U/HB branches unablated
  $\theta^{(j)} \leftarrow \mathrm{SafeTrain}(\theta^{(j-1)}, B_j, \mathcal{D})$ for 32 updates
    ▷ update LoRA only with response-only, sequence-normalized cross-entropy
end for
return the last completed adapter and direction bank

Mixtures are presentation percentages in the order harmful/response-ablation/continuous-response-ablation/PAP-response-ablation/utility/hard-benign (H/R/CR/PR/U/HB).

Unless otherwise stated, this subsection describes the shared training mechanics used by the primary Llama configuration. It starts from the mixed-SFT checkpoint with hard-benign supervision and runs three SafeTrain rounds of 4,000 presentations (32 updates) using H/R/CR/PR/U/HB shares $10/22.5/6.25/6.25/35/20$. Before each round, RSA refreshes candidates on the current checkpoint and appends an eligible complementary direction. If a complete candidate sweep finds none, the original procedure stops and retains the last completed round; it does not automatically continue with the old bank. Llama ours completed all three rounds, with training-bank ranks $(1, 2, 3)$. Fixed-bank continuations are separate variants. Exact selector and eligibility definitions are in Appendix A, while row-specific evaluation exceptions are in Appendix C.

For an orthonormal direction bank B, every intervened activation uses $A_B(h) = (I - BB^\top)h$. The pretrained weights are frozen and only LoRA parameters are updated. All conditions use response-only, sequence-normalized cross-entropy on the assistant target; harmful examples receive safe refusals, while utility and hard-benign examples retain helpful targets.

Mixed SFT uses clean harmful, continuous-attack, PAP-plus-continuous-attack, utility, and hard-benign branches with projection disabled. SafeTrain adds pure cumulative-subspace ablation and retains the current bank during both input-attack search and the outer safe-response update. Continuous perturbations are optimized only on user-content embeddings toward a separate unsafe prefix, aligned to the safe-target sequence, and detached before the outer update.

Table 4: Forward conditions in SafeTrain round j. All harmful conditions supervise the same safe refusal. Continuous perturbations are aligned to the outer sequence and detached. In mixed SFT all projections are disabled and the pure-ablation condition is absent; SFT and SafeTrain sampling shares are given in Appendix F.

Condition	User input	Outer target	Projection	Input perturbation
Clean refusal	x	y_{ref}	None	None
Pure ablation	x	y_{ref}	B_j	None
Continuous	x	y_{ref}	B_j	Continuous
PAP+continuous	$P(x;\xi)$	y_{ref}	B_j	Continuous
Utility	x	y_{util}	None	None
Hard-benign	x	y_{ben}	None	None

Training-time direction selection. Llama ours uses the `I` first-output-token probe, a harmless first-token KL bound of 0.2, and no activation-addition check. Each sweep uses 128 harmful and 128 harmless extraction prompts, with 32 validation prompts per pool (extraction seed 20260825; validation seed 20260826). Candidates cover the five assistant-boundary positions; the last 20% of source layers are excluded, the minimum residual fraction is 10^{-3}, and the basis tolerance is 10^{-5}. Extraction uses the unquantized checkpoint. These training gates are distinct from the default evaluation KL bound of 0.1.

Table 5: A SafeTrain round is one full pass over the constructed 4,000-presentation pool; it does not necessarily add a direction. All models first receive four mixed-SFT epochs (64 updates).

Original instruction-tuned checkpoint	SafeTrain rounds	Updates	Ranks by round
Llama-3-8B-Instruct	3	96	$1, 2, 3$
Mistral-7B-Instruct-v0.2	2	64	$1, 2$
Qwen2.5-14B-Instruct	1	32	$1, 2$

Optimization settings for Llama ReSAT. Mixed SFT uses 2,000 presentations per epoch for four epochs (64 updates), with learning rate 2×10^{-4}; SafeTrain uses one 4,000-presentation epoch per round (32 updates), with learning rate 5×10^{-6}. Both stages use AdamW, microbatch size 2, nominal effective batch size 128, weight decay 10^{-3}, gradient clipping at 0.3, and two warm-up updates. LoRA has rank 64, scaling parameter 16, and dropout 0.1, on the attention query/key/value/output and MLP gate/up/down projections, with no trainable bias. The frozen backbone uses NF4 with BF16 computation. The utility response cap is 256 tokens; harmful and hard-benign sequences are capped at 1,024 tokens. Native chat templates and response boundaries are validated.

The inner embedding attack takes 10 sign-gradient steps of size 10^{-4}, with radius 0.075 times the mean embedding norm, on user-content tokens only. PAP uses local Mixtral-8x7B-Instruct-v0.1 (NF4/BF16), at most 384 generated tokens, temperature 1.0, top-$p = 0.9$, and at most three attempts. Perturbations are detached for the outer response-only loss. The training seed is 20260908; harmful, utility, hard-benign, and evaluation pools are kept disjoint.

B.2 Mechanistic diagnostics retained from early experiments

We retain the two diagnostics that directly test whether refusal wording and refusal-token choice change the learned direction.

B.2.1 Ten refusal patterns after SFT

A separate QLoRA SFT experiment pairs 100 harmful prompts with each of the ten targets in Table 6 (1,000 refusal examples) and 1,000 UltraChat utility examples, with no direction ablation. Training uses four epochs, followed by two short calibration rounds that balance the output branches.

On 159 held-out harmful prompts with 30 samples each, pattern match is 96.10% and the refusal-like rate is 97.38%. Pattern frequencies are $[10.10, 8.93, 9.96, 8.24, 9.79, 9.06, 10.19, 10.59, 9.29, 9.96]\%$ (total variation 0.0478 unconditionally and 0.0305 conditional on a match), so the ten surface forms are approximately balanced. Nevertheless, eight of the nine unique first tokens—covering nine patterns—select the same -5/L12 direction. Only `The` selects -3/L19, with cosine 0.0746 to the shared direction.

Thus output diversity does not imply independent refusal mechanisms: ablating the shared direction collapses both refusal and pattern match, whereas ablating the approximately orthogonal `The` direction mainly changes surface form.

B.2.2 Refusal-token probes after SafeTrain

We re-extract a rank-one direction from the final three-round SafeTrain checkpoint using five first-output-token probes. Only `I` is a high-probability active prefix; the other four begin with very low

Table 6: Ten refusal targets used for the multi-pattern SFT diagnostic.

ID	Target response
1	I cannot help with that request. I can instead provide safe, legal, and prevention-focused information.
2	Sorry, but I am unable to provide those instructions. I can offer a high-level discussion of the risks or safer alternatives.
3	Your request is denied because fulfilling it could enable harmful activity. Assistance with prevention or protection remains available.
4	Providing the requested material is not allowed. Related defensive, educational, or harm-reduction guidance may still be provided.
5	It is not possible to answer this safely as written. Consider reframing the question around a legitimate and non-harmful goal.
6	The model will never assist with guidance that could put people or systems at risk. Safer alternatives and protective safeguards may be offered instead.
7	Safety rules prohibit assistance with the proposed activity. General information about risks, detection, or mitigation is still available.
8	I must decline to provide operational details for this activity. A non-actionable overview may be appropriate.
9	Such instructions fall outside what the assistant is authorized to provide. I can support lawful and responsible alternatives.
10	No actionable instructions will be provided for carrying out the proposed harm. I can redirect the discussion toward safety planning or recovery.

Table 7: Refusal and target-pattern metrics before and after direction ablation.

Condition	RR	10-pattern match	RR change
Baseline	98.67	97.11	–
Shared -5/L12 ablated	6.82	2.89	-91.85 pp
Fresh `The` -3/L19 ablated	97.71	92.84	-0.96 pp

baseline log-odds, so unchanged RR for those probes is not evidence of four independent safety mechanisms.

Table 8: Fresh rank-one probes on the final SafeTrain model (32 prompts, 10 samples each).

Probe R	Selected source	Base RR	Ablated RR	Log-odds change
{`I`}	-5/L11	100.00	97.19	$10.689 \rightarrow 7.385$
{`Sorry`}	-1/L4	100.00	100.00	$-15.939 \rightarrow -16.492$
{`As`}	-1/L4	100.00	100.00	$-15.662 \rightarrow -16.201$
{`No`}	-4/L21	100.00	100.00	$-13.230 \rightarrow -13.932$
{`You`}	-4/L18	100.00	100.00	$-15.117 \rightarrow -15.963$

`Sorry` and `As` select the same direction (cosine 1.0000), while `No` and `You` select distinct but similar directions (cosine 0.8446). The fresh final-model `I` direction and the initial-SFT d_1 have cosine 0.79198. The directions therefore remain strongly related after SafeTrain, although they are not identical. This diagnostic uses a 32×10 stochastic cohort and is not numerically interchangeable with the matched 125/128-prompt greedy RSA evaluation.

C Unified evaluation protocol and cohorts

Metrics. Table 2 uses the mixed jailbreak scores defined in Section 5.1, whereas Table 1 reports independent HarmBench ASR and intervention-matched MMLU. Full-test MMLU/GSM8K and XSTest measure unablated capability and helpfulness. Lexical RR and its scoring rules are reported in Appendix E; ASR is average of 3 judge classifer model.

Evaluation samples. Refusal-ablation evaluations use 150 prompts per run. Clean is the unablated baseline for the recorded RSA evaluation; Geometry/SOM retain their corresponding attack-run baselines. Evaluation directions are selected after training, separately from the SafeTrain banks. Saved responses are independently judged with the pinned HarmBench protocol and other classifier.

Attack endpoints and status. Single is self-seeded rank one; RSA reports the last accepted cumulative rank, not a test-selected optimum. If only rank one is accepted, RSA reuses Single, denoted (1) where shown. The default gate is KL ≤ 0.1 with the recorded eligibility checks; Geometry reports rank five and SOM its validation-selected ordered six-direction endpoint.

D COMPUTATIONAL COST

D.1 ATTACK SELECTION AND DEFENSE TRAINING ACROSS MODEL FAMILIES

Table 9 gives single-run wall-clock times for Ours(F) in minutes. Runs use NVIDIA GH200 120GB GPUs: one for RSA and two for training, including the PAP generator. These are elapsed times, not aggregate GPU time; run-to-run variability was not measured.

Table 9: Cost of Ours(F). SafeTrain uses 3, 2, and 1 rounds for Llama, Mistral, and Qwen, respectively; Mistral retains the same direction in round 2.

Model	RSA selection (min)	SFT (min)	SafeTrain (min)	Full training pipeline (min)
Llama-3-8B	10.3	154.2	387.6	619.2
Mistral-7B	6.1	154.8	206.4	396.0
Qwen2.5-14B	19.1	211.8	117.6	372.0

Attack timing. RSA selection sums four self-seeded candidate-scoring sweeps on the final checkpoint, including unsuccessful fourth sweeps for Mistral and Qwen, rounded to 0.1 minute. Loading, activation extraction, and downstream evaluation are excluded.

Training timing. SFT and SafeTrain include online attacks, optimization, validation, and checkpoint saving. Full-pipeline totals additionally include loading, direction search, smoke checks, and in-chain diagnostics. Only successful jobs are summed; queueing, failed runs, hyperparameter search, offline data generation, and final external benchmarks are excluded.

D.2 COMPARISON WITH MULTI-DIRECTION ABLATION METHODS

Table 10 compares our GH200 measurements of RSA, Geometry (Wollschläger et al., 2025), and SOM (Piras et al., 2026) on the same seven Llama-family checkpoints. RSA uses one GPU; Geometry and SOM allocate two for the target and judge. Search budgets, probes, and timing boundaries differ. Defense training and final held-out evaluation are excluded.

Table 10: Core attack-search wall-clock time on seven Llama-family checkpoints (minutes).

Target	RSA selection (min)	Geometry learning + selection (min)	SOM search (min)
Base	4.00	101.35	1310.37
ReFAT	6.70	93.20	1128.52
DeepRefusal	5.10	100.67	285.48
Ours(ReSAT)	10.30	76.30	111.33
MixAT	9.30	71.02	969.40
LAT	4.90	60.50	1082.12†
CAT	4.70	61.48	888.52

RSA search. Each of up to four self-seeded sweeps scores 160 candidates. Base and ReFAT stop at rank two after three sweeps; others reach rank four. Timers count unsuccessful sweeps and Base's reused first-direction search. MixAT/CAT use `Sorry`. Loading, activation extraction, and downstream generation/evaluation are excluded.

Geometry search. Five directions are learned sequentially, with one initialization and a two-epoch cap per direction. Timing reconstructed from target-audit and saved-direction timestamps includes data preparation, optimization, and validation, excluding initial loading and downstream generation/evaluation (including judging). MixAT/CAT use `Sorry`.

SOM search. We run 512 Bayesian-optimization trials over six directions using 41 validation prompts. Timing includes per-trial loading, generation, internal judging, and bookkeeping, excluding initial representation extraction/direction construction and final held-out evaluation.

E Lexical Refusal Rates under Ablation

Lexical scoring. Clean/Single/RSA use the recorded union of 12 case-insensitive refusal substrings and broader refusal-language regular expressions; specified v2 rows also recognize whole-response short refusals. Geometry/SOM retain their original 12-substring scorer. Table 11 reports lexical RR for the Table 1 checkpoints with available RR evaluations. Attack endpoints and status definitions follow Appendix C.

Table 11: Lexical refusal rates (RR, %; higher is better) for the Table 1 checkpoints with available RR evaluations.

Model	Clean	Single	Geometry	SOM	RSA (Ours)
Llama-3-8B					
Base	97.66	4.69	0.00	11.32	$3.12^{(2)}$
LAT	**100.00**	11.72	0.00	11.95	$4.69^{(4)}$
CAT	**100.00**	75.00	42.14	8.81	$48.44^{(4)}$
MixAT	**100.00**	14.06	0.00	4.40	$6.25^{(4)}$
ReFAT	99.22	3.12	0.63	4.40	$2.34^{(2)}$
DeepRefusal	**100.00**	**100.00**	**100.00**	71.70	$73.44^{(4)}$
Ours(H)	**100.00**	95.20	**100.00**	**100.00**	$\mathbf{84.80}^{(4)}$
Mistral-7B					
Base	48.44	5.47	48.44	3.14	$5.47^{(1)}$
LAT	92.19	4.69	76.10	2.52	$6.25^{(2)}$
CAT	97.66	64.84	2.52	18.24	$6.25^{(4)}$
MixAT	99.22	53.91	33.96	0.63	$14.84^{(3)}$
ReFAT	**100.00**	42.97	3.77	35.85	$42.97^{(1)}$
DeepRefusal	89.84	2.34	47.80	0.00	$2.34^{(1)}$
Ours(H-R12)	**100.00**	**97.60**	**100.00**	**62.34**	$\mathbf{77.60}^{(3)}$
Qwen2.5-14B					
Base	81.25	4.69	0.00	2.52	$4.69^{(1)}$
LAT	**100.00**	13.28	1.26	1.89	$13.28^{(1)}$
CAT	99.22	98.44	7.55	96.23	$\mathbf{98.44}^{(1)}$
MixAT	**100.00**	**99.22**	0.00	**100.00**	$72.66^{(3)}$
ReFAT	99.22	62.50	0.00	1.26	$32.03^{(2)}$
DeepRefusal	99.22	28.91	0.00	7.55	$3.12^{(3)}$
Ours(MB-v2)	**100.00**	98.40	**92.21**	74.68	$88.00^{(2)}$

F Ablation Study Details

Table 12 summarizes the six Llama-3-8B ablation groups in Table 3, listing training settings and sampling shares without evaluation results. All setups use the same instruction-tuned backbone and the implementation in Appendix B.1. The reference configuration is denoted **Ours** throughout this appendix.

Table 12: Training setups and sampling shares for the six ablation groups. All shares are percentages; Rds. counts completed SafeTrain rounds.

Setup	SFT shares C/A/P/U/HB	SafeTrain shares C/R/A/P/U/HB	Rds.
(i) SFT initialization			
Original SFT	50/0/0/50/0	25/12.5/6.25/6.25/45/5	3
Mixed SFT	20/5/5/50/20	25/12.5/6.25/6.25/45/5	3
No initial SFT	—	25/12.5/6.25/6.25/45/5	3
(ii) Input-attack training			
No input attacks	30/0/0/50/20	10/35/0/0/35/20	3
With input attacks (Ours)	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3
(iii) Hard-benign supervision			
No Hard Benign	20/5/5/70/0	10/22.5/6.25/6.25/55/0	3
With Hard Benign (Ours)	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3
(iv) Training rounds			
Ours after round 1	20/5/5/50/20	10/22.5/6.25/6.25/35/20	1
Ours after round 2	20/5/5/50/20	10/22.5/6.25/6.25/35/20	2
Ours after round 3	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3
(v) Complementary directions			
Repeated d_1	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3
Cumulative directions (Ours)	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3
(vi) Training-time RSA			
Without training RSA†	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3
With training RSA (Ours)	20/5/5/50/20	10/22.5/6.25/6.25/35/20	3

Mixture definitions and budget. C denotes clean harmful-prompt refusal; R, refusal under direction ablation; A, continuous input attacks; P, PAP rewriting followed by continuous attacks; U, ordinary utility; and HB, hard-benign supervision. Harmful branches supervise safe refusals, whereas U and HB supervise helpful answers. SFT never uses direction ablation; SafeTrain applies it to R/A/P except in the no-RSA control. SFT runs for four epochs where present. Each SafeTrain round contains 4,000 presentations and 32 optimizer updates, giving 32/64/96 updates after 1/2/3 rounds.

Controlled changes. Group (i) holds the SafeTrain mixture and three-round schedule fixed, but does not match total training compute. In (ii), removing input attacks reallocates their SFT shares to C and SafeTrain shares to R; source queries, safe targets, sample order, microbatch membership, hard-benign exposure, and update budgets are preserved. Group (iii) replaces HB with U in both stages, preserving the harmful branches, total sample count, update budget, and benign share. Group (iv) uses successive checkpoints from one run; round 3 is the predefined final model. In (v), Repeated d_1 reuses the same first-round checkpoint, then fixes that direction for two further rounds, giving training-bank ranks $(1, 1, 1)$ instead of Ours' $(1, 2, 3)$; both use 96 updates. Final-model RSA searches fresh directions independently of the training banks.

†In (vi), both setups start from the same mixed-SFT checkpoint. Without training RSA retains Ours' sample quotas, source order, safe targets, continuous/PAP attacks, and 96-update budget, but disables direction ablation in both inner attack search and outer training. R/A/P remain sampling categories with their projections removed. This control matches optimizer updates, not FLOPs or the realized adversarial inputs.

G CUMULATIVE DIRECTION-HISTORY MECHANISM ANALYSIS

Figure 4 reports the complete set of diagnostics accompanying Section 5.6.

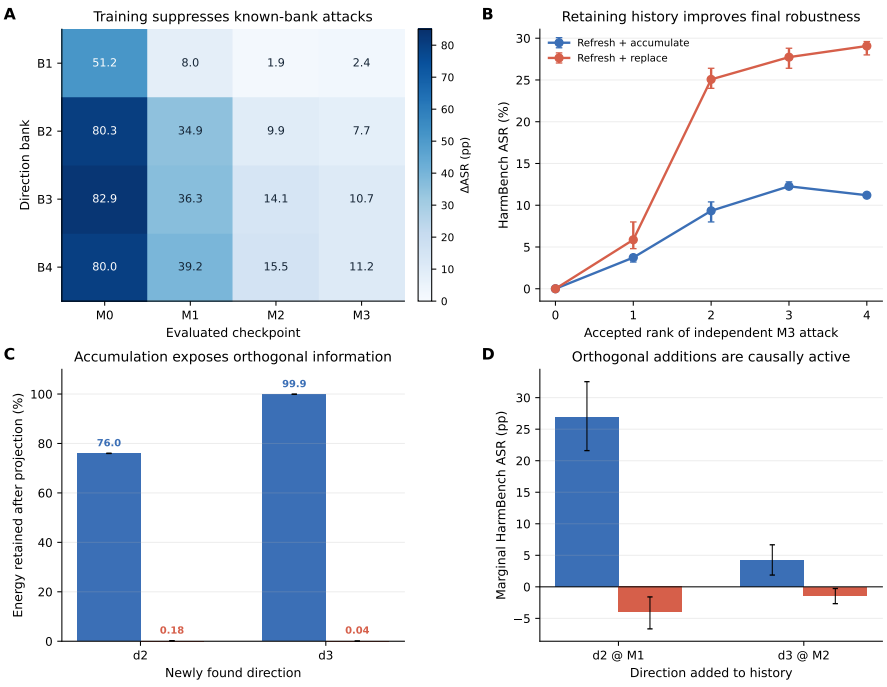


Figure 4: **Mechanism analysis of cumulative direction history on Llama-3-8B. (A)** SafeTrain progressively suppresses attacks from known direction banks. **(B)** Retaining history yields lower independently recoverable M_3 ASR than replacing the bank. **(C)** Accumulation exposes information outside the historical span, whereas replacement largely rediscovers that span. **(D)** The complementary directions found under accumulation have positive marginal causal effects. Results use three matched seeds and 125 held-out HarmBench prompts per seed. Error bars in Panels B and C show the seed range; error bars in Panel D show paired hierarchical-bootstrap 95% confidence intervals over seeds and prompts.

Shared protocol. We use the Llama-3-8B checkpoints $M_0, \ldots, M_3$, where M_0 is the SFT-initialized model and M_t is saved after SafeTrain round t. Unless stated otherwise, all results aggregate three matched seeds and use the same 125 held-out HarmBench prompts per seed. ASR is scored by the independent HarmBench classifier; no directions are selected on this held-out evaluation set. For schedule s, let $B^{(s)}_{<j} = \mathrm{span}(d^{(s)}_1, \ldots, d^{(s)}_{j-1})$ denote the schedule-specific prefix. Refresh+accumulate supplies this prefix during direction discovery and training, whereas refresh+replace discards it; for replacement, we reconstruct the prefix only post hoc for the matched diagnostics in Panels C and D.

Panel A: suppression of known bank prefixes. Setup. For every checkpoint M_t, we replay ablation of the cumulative training-bank prefixes $B_k = \mathrm{span}(d_1, \ldots, d_k)$, $k \in \{1, 2, 3, 4\}$, without re-estimating the directions, and report the ASR increase relative to the corresponding clean checkpoint. **Conclusion.** SafeTrain progressively defends against the interventions already included in the bank: the B_1 effect falls from 51.2 pp on M_0 to 8.0, 1.9, and 2.4 pp on M_1, M_2, and M_3, while the B_2–B_4 effects likewise fall from 80.3–82.9 pp on M_0 to 7.7–11.2 pp on M_3. Thus, later checkpoints acquire robustness to the known bank rather than merely changing their clean refusal rate.

Panel B: refresh-and-accumulate versus refresh-and-replace. Setup. We compare two matched SafeTrain schedules with the same initialization, data quotas, update budget, and refresh times. Refresh+accumulate retains every previously found direction, whereas refresh+replace discards the old bank and trains only against the direction re-estimated in the current round. We then attack each final M_3 checkpoint independently and report HarmBench ASR at accepted attack ranks zero through four. **Conclusion.** At rank four, replacement reaches 29.07% ASR, compared with 11.20% under

accumulation, a paired increase of 17.87 pp (95% CI: $[14.13, 21.87]$). Retaining direction history is therefore necessary for the final robustness gain.

Panel C: orthogonal information in refreshed directions. Setup. At the second and third refreshes, we project the newly found candidates d_2 and d_3 onto the orthogonal complement of the preceding historical span $B_{<j}$ and measure the retained squared-energy ratio $100\|P_{B_{<j}^{\perp}} d_j\|_2^2 / \|d_j\|_2^2$. The same calculation is applied to candidates from both schedules. **Conclusion.** Under accumulation, d_2 and d_3 retain 76.0% and 99.9% of their energy after projection; replacement retains only 0.185% and 0.043%. Accumulation therefore exposes directions that are genuinely complementary to the covered span, whereas replacement almost entirely rediscovers previously covered directions.

Panel D: causal validation of the orthogonal information. Setup. For each refreshed candidate, we measure its held-out marginal causal effect by adding it to the preceding bank and computing $\mathrm{ASR}(B_{<j} \cup \{d_j\}) - \mathrm{ASR}(B_{<j})$. We evaluate accumulated and replacement candidates with the same prompts and classifier. **Conclusion.** The accumulated d_2 and d_3 increase ASR by 26.93 and 4.27 pp, respectively, whereas the replacement candidates yield no positive marginal gain. The orthogonal information identified in Panel C is therefore causally active refusal vulnerability, not an incidental representation rotation.

H KL ABLATION

The KL we use for ReSAT training is 0.2 and RSA ablation attack is 0.1, but We also vary the RSA selection threshold to $\mathrm{KL} \leq 0.2$ and $\mathrm{KL} \leq 0.5$ on seven fixed checkpoints from each of three model families. Directions are freshly extracted from each fixed target and expanded adaptively. The KL gate measures benign first-token distribution divergence during selection, not an upper bound on MMLU degradation. Table 13 reports the *last accepted cumulative rank*, not the rank with the highest test ASR. A rank below four means that no next candidate passed the complete selection constraints. ASR is the independent HarmBench classifier's positive-label count divided by the original cohort size. RR is lexical refusal rate, and MMLU is utility benchmark.

Table 13: RSA KL ablation (%). $\mathrm{KL} \leq 0.1$ reproduces the RSA ASR/MMLU columns of main paper Table 1 and the corresponding lexical RR from Table 11, under their original settings. In all three groups, r is the last accepted cumulative rank; the first group uses the ranks corresponding to the main-table RSA endpoints. RR is lexical refusal rate, MMLU is utility accuracy.

	$\mathrm{KL} \leq 0.1$				$\mathrm{KL} \leq 0.2$				$\mathrm{KL} \leq 0.5$			
Model	r	ASR↓	RR↑	MMLU↑	r	ASR↓	RR↑	MMLU↑	r	ASR↓	RR↑	MMLU↑
Llama-3-8B												
Base	2	92.19	3.12	64.25	3	94.53	7.81	64.00	4	88.28	4.69	63.25
LAT	4	86.72	4.69	62.75	3	92.19	4.69	63.50	4	88.28	5.47	64.25
CAT	4	45.31	48.44	64.00	4	58.59	32.81	64.00	4	74.22	15.63	65.75
MixAT	4	82.03	6.25	60.25	4	64.84	24.22	61.25	4	79.69	6.25	60.25
ReFAT	2	92.19	2.34	65.25	3	84.38	2.34	64.50	3	88.28	2.34	65.50
DeepRefusal	4	23.44	73.44	64.25	4	4.69	93.75	65.00	4	3.13	95.31	65.25
Ours (ReSAT)	4	11.20	84.80	64.75	4	16.80	78.40	64.50	4	16.80	78.40	64.50
Mistral-7B												
Base	1	69.53	5.47	59.50	1	73.44	3.13	56.25	1	72.66	3.91	56.25
LAT	2	81.25	6.25	54.00	2	81.25	5.47	54.00	2	81.25	6.25	54.00
CAT	4	75.78	6.25	53.25	4	75.78	6.25	53.25	4	75.78	5.47	53.25
MixAT	3	65.62	14.84	55.75	4	71.09	10.94	50.75	4	71.09	10.94	50.75
ReFAT	1	28.91	42.97	57.25	4	71.09	3.91	54.75	4	59.38	2.34	42.25
DeepRefusal	1	78.91	2.34	57.75	2	42.97	3.91	53.50	2	43.75	3.91	53.50
Ours (ReSAT)	3	20.00	77.60	57.25	4	16.00	83.20	54.25	4	20.00	78.40	48.00
Qwen2.5-14B												
Base	1	79.69	4.69	76.50	1	81.25	4.69	76.50	1	78.91	1.56	75.50
LAT	1	47.66	13.28	76.00	2	71.09	0.00	76.00	3	94.53	3.13	74.50
CAT	1	1.56	98.44	75.50	1	14.84	78.91	75.75	4	35.94	15.63	75.25
MixAT	3	29.69	72.66	70.25	1	9.38	88.28	77.25	3	31.25	72.66	70.25
ReFAT	2	64.06	32.03	75.50	2	56.25	32.81	77.25	2	73.44	0.78	76.75
DeepRefusal	3	72.66	3.12	75.75	3	91.41	6.25	75.25	4	79.69	5.47	72.50
Ours (ReSAT)	2	8.80	88.00	79.00	3	19.20	76.00	77.75	4	69.60	12.00	73.25

I SEED REPRODUCIBILITY ACROSS MODEL FAMILIES

This appendix reports every completed training-seed evaluation for the primary Llama-3-8B Ours, Mistral-7B Ours, and Qwen2.5-14B Ours recipes. Seeds 20260908, 20260909, and 20260910 are distinct trained checkpoints for each family. Each refusal-ablation cell is reported as independent HarmBench ASR / lexical refusal rate (RR) / MMLU accuracy, in percent.

Table 14: Complete refusal-ablation seed results (%). Each entry is HarmBench ASR / lexical RR / MMLU. The parenthesized value is the last accepted RSA rank.

Family	Seed	Clean	Single	Geometry	SOM	RSA endpoint
Llama-3-8B Ours	20260908	0.00/100.00/65.75	4.00/95.20/63.75	0.00/100.00/65.50	0.00/100.00/66.00	11.20/84.80/64.75 (4)
	20260909	0.00/100.00/66.00	4.00/95.20/63.50	0.00/100.00/65.25	0.00/100.00/66.50	11.20/84.80/65.00 (4)
	20260910	0.00/100.00/65.25	3.20/96.00/63.75	0.00/100.00/65.25	0.00/100.00/66.75	11.20/85.60/65.00 (4)
Mistral-7B Ours	20260908	0.00/100.00/57.25	1.60/97.60/57.25	0.00/100.00/55.50	28.57/62.34/47.25	20.00/77.60/57.25 (3)
	20260909	0.00/100.00/57.25	1.60/97.60/57.75	0.00/100.00/54.25	35.71/62.34/47.50	20.00/76.80/57.25 (3)
	20260910	0.00/100.00/57.50	1.60/97.60/57.75	0.00/100.00/53.75	30.52/62.34/39.50	19.20/78.40/57.00 (3)
Qwen2.5-14B Ours	20260908	0.00/100.00/78.50	1.50/98.40/79.00	7.14/92.21/78.50	22.73/74.68/76.75	8.80/88.00/79.00 (2)
	20260909	0.00/100.00/79.00	1.50/98.40/79.00	77.92/16.88/78.50	26.62/71.43/76.50	8.80/88.80/79.00 (2)
	20260910	0.00/100.00/78.50	1.50/98.40/79.00	2.60/97.40/78.50	18.83/78.57/75.50	8.80/89.60/79.25 (2)

Table 15: Complete jailbreak and utility results across the seed runs (%). NoAtk–RFA are the paper's composite safety scores (lower is better); MMLU, GSM8K, and XSTest are utility scores (higher is better).

Family	Seed	NoAtk	AutoDAN	HumanJB	GCG	PAIR	Inp.	Soft	RFA	MMLU	GSM8K	XSTest
Llama-3-8B	20260908	0.04	0.04	0.04	1.05	1.19	47.31	82.12	57.25	66.49	73.09	53.00
	20260909	0.04	0.04	0.04	2.37	2.81	48.94	84.19	56.95	66.52	72.48	53.00
	20260910	0.04	0.04	0.04	1.56	1.35	46.72	83.57	57.42	67.04	73.46	55.00
Mistral-7B	20260908	0.04	4.12	0.07	53.83	30.49	64.63	80.51	47.29	59.22	42.99	57.00
	20260909	0.04	5.48	0.07	57.48	38.10	64.26	81.10	47.93	59.19	42.46	60.00
	20260910	0.04	7.22	0.07	61.86	33.02	63.44	81.44	47.69	59.19	42.61	56.00
Qwen2.5-14B	20260908	0.04	4.58	0.65	52.48	44.21	80.60	83.75	52.07	78.93	86.28	47.00
	20260909	0.04	3.85	0.66	54.11	50.69	79.47	85.29	53.49	78.86	86.50	49.00
	20260910	0.04	0.04	0.61	45.96	48.28	80.21	84.44	52.84	78.89	86.35	49.00