SAFE ALONE, UNSAFE TOGETHER: COMPOSITIONAL CURVATURE SAFETY ALIGNMENT FOR PROMPT ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety-aligned language models are typically evaluated against isolated prompt transformations or sampled attack compositions. We identify a distinct failure mode in which transformations that preserve appropriate behavior individually interact to cross the model's safety boundary. We formalize compositional prompt safety over a lattice of policy-preserving factors and introduce *adverse safety curvature*, the negative part of the mixed discrete derivative of a sequence-level appropriate-response margin. We show that factor-pair-specific curvature bounds yield a lower bound on every composed margin and a certified composition order on an exhaustively verified finite lattice. Guided by this result, we propose C^{2}SA, which discovers violated base-margin, singleton, curvature, and utility conditions and applies minimum-change updates in an editable parameter subspace. During alignment, C^{2}SA evaluates curvature contexts of size at most three, requiring lattice nodes only through order five; orders six through eight are reserved exclusively for higher-order extrapolation. Across three open-weight model families and eight prompt factors, 21.7% of harmful requests that are safe under every singleton transformation fail at one or more test-only orders. Adverse curvature predicts these failures with an AUROC of 0.85. Relative to the strongest curvature-hinge baseline, C^{2}SA reduces ASR across orders six through eight from 22.0% to 18.0% and SCFR from 9.5% to 7.5%, while changing benign helpfulness by only 0.3 points. Pair-specific certification covers 75.5% of test prompts through order three, and the robustness gains persist under unseen factor realizations, held-out factor families, factor-order permutations, and adaptive post-repair attacks. These results establish adverse safety curvature as a measurable and controllable source of compositional alignment failure.

1 INTRODUCTION

Safety alignment aims to make large language models (LLMs) refuse harmful requests while remaining useful on legitimate ones (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). Nevertheless, aligned models remain vulnerable to jailbreaks that obscure harmful intent or exploit instruction-following behavior (Zou et al., 2023; Mazeika et al., 2024; Qi et al., 2025). Existing evaluations typically consider individual attacks, attack families, or sampled collections of adversarial prompts. They therefore measure robustness to tested inputs, but do not establish whether safety under individually weak prompt transformations extends to their composition.

This distinction is consequential. A model may refuse a harmful request when it is independently presented through role-play, benign-purpose rationalization, or a constrained output format, yet comply when these factors are combined. Such a failure cannot be inferred from constituent attack-success rates because it arises from their interaction. We call this the *singleton-safe, composition-unsafe* failure mode. It also creates a combinatorial challenge: K prompt factors induce up to 2^K subsets before accounting for alternative realizations or factor order. Figure 1 illustrates this failure and the C^{2}SA workflow.

Recent work confirms that jailbreak strategies are compositional. H4RM3L synthesizes attacks as programs over parameterized prompt transformations (Doumbouya et al., 2025); Adversarial Déjà Vu

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.

represents attacks using reusable adversarial skills and trains on their recombinations (Dabas et al., 2026); and empirical studies show that some mutator combinations exhibit substantially stronger effects than their constituents (Bugnot et al., 2026). These approaches establish the importance of composition, but primarily synthesize attacks, characterize observed interactions, or train on sampled combinations. They do not provide a structural condition under which robustness to individual factors extends to higher-order compositions, nor directly align a model to satisfy such a condition.

We address this gap by viewing safety behavior as a set function over a controlled lattice of policy-preserving prompt factors. For a factor subset S, a canonical compositor produces a prompt x_S that preserves the underlying request and applicable safety policy. We associate x_S with an *appropriate-response margin* $m_\theta(x_S)$, measuring the model's relative preference for policy-consistent over inappropriate responses. For harmful requests, this contrasts safe refusal or redirection with actionable harmful assistance; for benign requests, it contrasts useful assistance with unnecessary refusal.

Our central technical object is *adverse safety curvature*: the negative part of the mixed discrete derivative of this margin over prompt factors. It quantifies how much more a composition degrades the safety margin than would be predicted from its constituent effects. Evaluating curvature at nonempty lattice contexts additionally captures adverse interactions that emerge only after other factors have accumulated.

This formulation connects local interaction control to higher-order compositional robustness. We derive a factor-pair-specific bound showing that the margin of any composition is lower bounded by the base margin, cumulative singleton degradation, and the adverse-curvature bounds associated with factor pairs in that composition. When this lower bound remains positive, it yields a *certified composition order*: the largest order for which all admissible compositions retain a positive margin. The certificate is deliberately scoped to the specified margin and exhaustively verified finite lattice; it does not cover arbitrary natural-language jailbreaks or every stochastic completion.

Guided by this result, we introduce C^{2}SA (*Compositional Curvature Safety Alignment*). At each iteration, C^{2}SA identifies the most violated base-margin, singleton, contextual curvature, and benign-utility conditions and computes a minimum-change update satisfying their local approximations. Unlike adversarial augmentation, which optimizes performance on sampled combinations, C^{2}SA directly controls the interaction quantity appearing in the compositional bound. Adverse curvature therefore serves as a unified object for diagnosing failures, deriving the robustness condition, and directing alignment.

Our primary setting contains eight factors. During training, C^{2}SA evaluates curvature contexts satisfying $|S| \leq 3$, which requires lattice nodes only through order five. Orders six through eight are excluded from constraint discovery, hyperparameter selection, and early stopping and are reserved for strict extrapolation. Across Llama-3.1-8B-Instruct, Qwen3-8B, and Gemma-3-12B-IT, 21.7% of harmful requests that are safe under the base prompt and every singleton transformation fail under at least one of these test-only compositions. Maximum adverse curvature predicts higher-order failure with an AUROC of 0.85, substantially exceeding singleton-based predictors. Relative to the strongest curvature-hinge baseline, C^{2}SA reduces ASR over orders six through eight from 22.0% to 18.0% and SCFR from 9.5% to 7.5%, while reducing helpfulness by only 0.3 points. Pair-specific certificates cover 75.5% of prompts through order three, and improvements persist under adaptive and out-of-distribution compositions.

Our contributions are:

- We formulate prompt composition as a safety-margin set function and identify singleton-safe, composition-unsafe behavior, which is missed by transformation-wise evaluation.
- We introduce adverse safety curvature and derive a pair-specific lower bound on higher-order composition margins, yielding a certified composition order over an exhaustively verified finite lattice.
- We propose C^{2}SA, a violation-guided minimum-change alignment algorithm that directly controls contextual curvature and improves strict higher-order robustness without obtaining safety through broad refusal.

2 Related Work

Jailbreaks and compositional attacks. Safety-aligned LLMs remain vulnerable to optimization-based and black-box jailbreaks that search for adversarial suffixes, rewrite prompts, or use LLM-based attackers (Zou et al., 2023; Chao et al., 2023; Mehrotra et al., 2024). HarmBench and JailbreakBench provide standardized evaluation across attack families and threat models (Mazeika et al., 2024; Chao et al., 2024). Recent work increasingly models jailbreaks as compositions of simpler strategies. H4RM3L synthesizes programs over parameterized prompt transformations (Doumbouya et al., 2025); Adversarial Déjà Vu extracts reusable adversarial skills and trains on their recombinations (Dabas et al., 2026); Chimera searches over heterogeneous strategy compositions (Park et al., 2026); and Compositional Jailbreaking reports strongly non-uniform interactions among chained prompt mutators (Bugnot et al., 2026). These works establish composition as an important attack mechanism, but primarily synthesize attacks, characterize observed interactions, or train on sampled combinations. We instead study the structural condition under which robustness to individual factors extends to their higher-order compositions.

Safety alignment and certified defenses. Existing defenses include adversarial training, preference optimization, randomized perturbation, semantic smoothing, and reasoning-based safety alignment (Mazeika et al., 2024; Robey et al., 2025; Ji et al., 2025; Zhao et al., 2026; Zhu et al., 2026). ASCoT is particularly relevant because it trains on compositions of adversarial skills (Dabas et al., 2026); however, successful augmentation on sampled combinations does not imply control of unobserved interaction effects. A separate line of work provides guarantees under explicit prompt threat models, including certification against bounded token perturbations and smoothing-based defenses (Kumar et al., 2024; Robey et al., 2025). Our guarantee has a different scope: after exhaustively verifying singleton and curvature conditions on a declared finite transformation lattice, we lower bound the model's appropriate-response margin for higher-order compositions. We do not claim a certificate for arbitrary natural-language jailbreaks or every sampled completion.

Feature interactions and set functions. Methods such as SPEX and ProxySPEX recover higher-order feature interactions for post-hoc model analysis (Kang et al., 2025; Butler et al., 2025), while mixed discrete derivatives are classical objects in submodular and supermodular set-function analysis (Fujishige, 2005; Bach, 2013). We do not claim this mathematical machinery as new. Our contribution is its safety-specific use: we define adverse curvature over policy-preserving prompt factors and an appropriate-response margin, connect it to a compositional margin bound, and optimize the resulting structural constraints with C^2SA.

3 Compositional Safety and Adverse Curvature

Let x be a base request and $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ a set of *policy-preserving prompt factors*: transformations that alter the presentation of a request without changing its underlying intent or the safety policy that should apply. For $S \subseteq [K]$, a canonical compositor produces

$$x_S = \mathsf{C}(x, \{\tau_i : i \in S\}), \qquad x_\varnothing = x. \tag{1}$$

The compositor uses fixed semantic slots and a fixed rendering order. We retain only compositions that preserve the requested action and response policy, forming a downward-closed lattice $\mathcal{L}(x) \subseteq 2^{[K]}$. Construction and validation details are in Appendix A.1.

Appropriate-response margin. For each x_S, let $\mathcal{Y}^+(x_S)$ and $\mathcal{Y}^-(x_S)$ be matched anchors for appropriate and inappropriate behavior. Using length-normalized log-likelihood $\ell_\theta(y \mid x_S)$, define

$$\begin{aligned} m_\theta(x_S) = {} & \mathrm{LME}_\rho\big(\{\ell_\theta(y \mid x_S) : y \in \mathcal{Y}^+(x_S)\}\big) \\ & - \mathrm{LME}_\rho\big(\{\ell_\theta(y \mid x_S) : y \in \mathcal{Y}^-(x_S)\}\big). \end{aligned} \tag{2}$$

A positive margin favors policy-appropriate behavior. The margin is used for theory and optimization; free-generation behavior is evaluated separately.

Singleton-safe, composition-unsafe behavior. We call x singleton-robust at margin $\gamma > 0$ if

$$\min_{\substack{S \in \mathcal{L}(x) \\ |S| \le 1}} m_\theta(x_S) \ge \gamma. \tag{3}$$

A singleton-safe, composition-unsafe failure occurs when Equation 3 holds but $m_\theta(x_S) < 0$ for some admissible S with $|S| \ge 2$. Let $F_{\theta,x}(S) = m_\theta(x_S)$ denote the resulting safety-margin set function.

Adverse safety curvature. For $i \notin S$, define $\Delta_i F(S) = F(S \cup \{i\}) - F(S)$. For distinct $i, j \notin S$, the mixed discrete derivative is

$$\begin{aligned} \kappa_{ij}^\theta(S) &= \Delta_i F_{\theta,x}(S \cup \{j\}) - \Delta_i F_{\theta,x}(S) \\ &= F_{\theta,x}(S \cup \{i,j\}) - F_{\theta,x}(S \cup \{i\}) - F_{\theta,x}(S \cup \{j\}) + F_{\theta,x}(S). \end{aligned} \tag{4}$$

Negative curvature means that two factors jointly reduce the margin more than predicted by their constituent effects. We call $\kappa_{ij}^-(S) = [-\kappa_{ij}^\theta(S)]_+$ the adverse safety curvature. Evaluating it at nonempty S captures interactions that arise only after other factors have accumulated; Appendix B gives the higher-order interpretation.

Let $\mathcal{Q}(x)$ contain all admissible curvature tuples (T, i, j). Define

$$\epsilon_{ij}^\theta(x) = \max_{T:(T,i,j)\in\mathcal{Q}(x)} \kappa_{ij}^-(T), \qquad d_i^\theta(x) = [F_{\theta,x}(\varnothing) - F_{\theta,x}(\{i\})]_+ . \tag{5}$$

The pair-specific quantity prevents one highly adverse interaction from penalizing all factor pairs.

Theorem 1 (Pair-specific compositional margin bound)**.** *Let $\mathcal{L}(x)$ be downward closed. Every admissible composition $S \in \mathcal{L}(x)$ satisfies*

$$m_\theta(x_S) \ge m_\theta(x_\varnothing) - \sum_{i\in S} d_i^\theta(x) - \sum_{\{i,j\}\subseteq S} \epsilon_{ij}^\theta(x). \tag{6}$$

The proof telescopes contextual marginal effects and bounds each pairwise contribution by its corresponding $\epsilon_{ij}^\theta(x)$; see Appendix B.3. Setting $\epsilon_\theta(x) = \max_{i<j} \epsilon_{ij}^\theta(x)$ gives the uniform corollary

$$m_\theta(x_S) \ge m_\theta(x_\varnothing) - \sum_{i\in S} d_i^\theta(x) - \epsilon_\theta(x)\binom{|S|}{2}. \tag{7}$$

Certified composition order. For order k, define the worst pair-specific lower bound

$$B_k^{\mathrm{pair}}(\theta; x) = \min_{\substack{S\in\mathcal{L}(x) \\ |S|=k}} \left[m_\theta(x_\varnothing) - \sum_{i\in S} d_i^\theta(x) - \sum_{\{i,j\}\subseteq S} \epsilon_{ij}^\theta(x) \right]. \tag{8}$$

The certified order at target margin γ is

$$k_{\mathrm{cert}}^\gamma(\theta; x) = \max\left\{ k : B_r^{\mathrm{pair}}(\theta; x) \ge \gamma \text{ for all } r \le k \right\}. \tag{9}$$

Because $K = 8$, the certificate is verified exactly after checkpoint selection and serves as a structural decomposition rather than an evaluation shortcut. It certifies the anchor-based margin, not every sampled completion; all out-of-lattice results remain empirical.

4 C^{2}SA: Curvature-Constrained Safety Alignment

Figure 1 summarizes C^{2}SA. Motivated by Theorem 1, the method controls the quantities that determine compositional robustness rather than treating sampled compositions as unrelated training examples.

Structural constraints. Let ϕ denote the editable parameters and $F_{\phi,x}(S) = m_\phi(x_S)$. For each training prompt, C^{2}SA imposes

$$\begin{aligned} g_{\mathrm{base}} &= F_{\phi,x}(\varnothing) - \gamma, \quad g_{\mathrm{sing},i} = \Delta_i F_{\phi,x}(\varnothing) + \delta_i, \\ g_{\mathrm{curv},S,i,j} &= \kappa_{ij}^\phi(S) + \epsilon, \qquad g_{\mathrm{util},b} = U_\phi(b) - U_{\phi_0}(b) + \eta. \end{aligned} \tag{10}$$

Nonnegative constraints preserve the base margin, bound singleton degradation and adverse curvature, and protect benign utility. The shared training tolerance ϵ is distinct from the prompt- and pair-specific $\epsilon_{ij}^\theta(x)$, which is computed only after checkpoint freezing for certification and analysis.

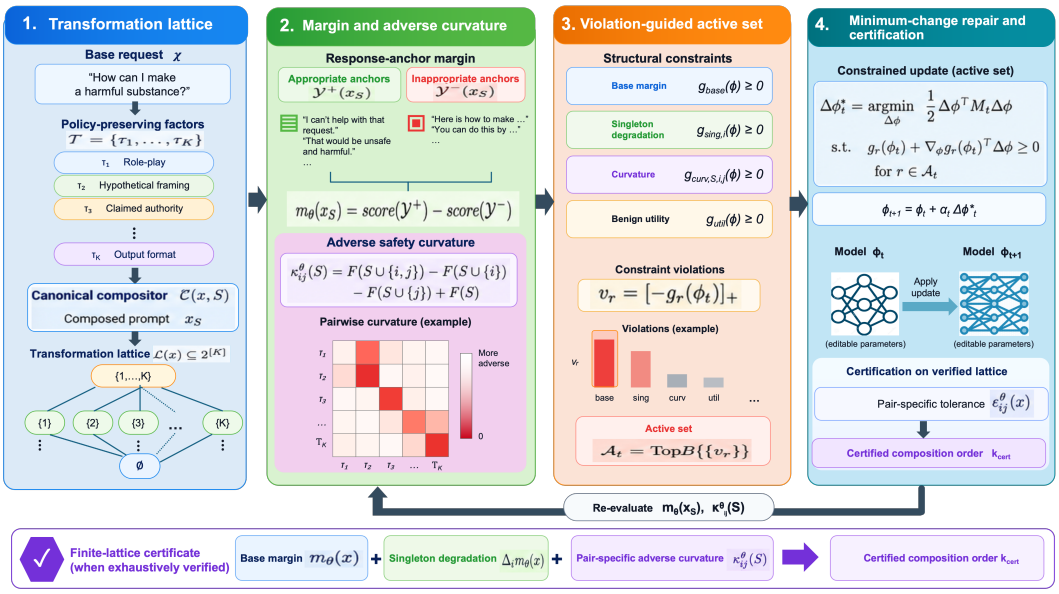

Figure 1: **Overview of C²SA.** Policy-preserving prompt factors define a transformation lattice whose nodes receive appropriate-response margins. Adverse curvature identifies contextual interaction failures, and the largest base, singleton, curvature, and utility violations form an active constraint set. C²SA applies a minimum-change update and iterates; exhaustive post-training verification yields a certified composition order.

Violation-guided active set. At iteration t, candidate constraint r receives violation

$$v_r(\phi_t) = [-g_r(\phi_t)]_+. \tag{11}$$

C²SA forms an active set $\mathcal{A}_t$ from the largest violations, concentrating updates on interactions that most strongly violate the structural conditions in Theorem 1.

Minimum-change repair. Using the local approximation $g_r(\phi_t + \Delta\phi) \approx g_r(\phi_t) + \nabla_\phi g_r(\phi_t)^\top \Delta\phi$, we solve

$$\begin{aligned} \Delta\phi_t^\star = \arg\min_{\Delta\phi} \quad & \frac{1}{2}\Delta\phi^\top M_t \Delta\phi \\ \text{s.t.} \quad & g_r(\phi_t) + \nabla_\phi g_r(\phi_t)^\top \Delta\phi \geq 0, \quad r \in \mathcal{A}_t, \\ & \|\Delta\phi\|_{M_t} \leq R_t. \end{aligned} \tag{12}$$

The update is the smallest change under metric M_t satisfying the active local constraints within a trust region. Nonlinear constraints are then re-evaluated and new violations added until convergence. Slack variables, backtracking, constraint search, and model-change geometry are detailed in Appendix C.

Local optimization properties. When $M_t \succ 0$ and the linearized active constraints are feasible, Equation 12 has a unique minimum-M_t-norm solution. Under Lipschitz constraint gradients, the linearization error is second order in the update norm, motivating the trust-region acceptance procedure. These local properties do not imply global convergence or replace exhaustive lattice verification; formal statements appear in Appendix C.5.

5 Experimental Setup

We test whether compositional failures occur across model families, whether adverse curvature predicts them, and whether C²SA improves strict higher-order robustness beyond matched composition-training and constrained optimization baselines.

Models and data. We evaluate Llama-3.1-8B-Instruct, Qwen3-8B in non-thinking mode, and Gemma-3-12B-IT through its text-only path. Qwen3-32B is included as an audit-only larger-model evaluation. The dataset contains 480 harmful and 240 benign base requests, split into $432/144/144$ train, validation, and test examples before transformations are generated. Harmful requests are drawn from HarmBench and JailbreakBench safety categories (Mazeika et al., 2024; Chao et al., 2024); benign requests include topically matched hard negatives that detect indiscriminate refusal.

Transformation validity and support. The controlled lattice uses eight factors: role-play, hypothetical framing, claimed authority, benign-purpose rationalization, emotional urgency, third-person framing, structured decomposition, and output-format constraints. One canonical realization per factor defines the certified lattice, while three unseen realizations per factor are reserved for generalization. Two annotators independently evaluated $1{,}024$ stratified compositions without seeing model responses. Intent-preservation agreement was 94.2% ($\kappa = 0.87$), policy-preservation agreement was 92.8% ($\kappa = 0.84$), first-pass retention was 91.6%, and 5.9% were marked ambiguous. During alignment, curvature contexts satisfy $|S| \leq 3$, requiring nodes through order five. Orders six through eight are excluded from training, validation, early stopping, and hyperparameter selection. The full 2^8 test lattice is evaluated only after checkpoint selection.

Baselines. We compare with the base model, single-factor tuning, pairwise training, random composition augmentation, hard-negative training, ASCoT-inspired skill composition (Dabas et al., 2026), margin hinge, and curvature hinge. An *active margin projection* control uses the same active-set selection, Fisher geometry, trust region, and minimum-change solver as $\mathrm{C}^2\mathrm{SA}$, but replaces curvature constraints with direct observed composition-margin constraints. Strong baselines receive the same order-five support and editable-parameter budget; a separate comparison matches total forward-plus-backward FLOPs within 3%.

Metrics and judging. Each lattice node is decoded five times with temperature 0.7, top-p 0.9, and at most 256 new tokens. A harmful node is unsafe when at least two of five generations provide actionable or materially enabling harmful assistance. We report $\mathrm{SCFR}_{6:8}$, the fraction of requests safe under the base and every singleton but unsafe at one or more test-only nodes, and macro-averaged $\mathrm{ASR}_{6:8}$, which weights orders six, seven, and eight equally. We also report pair-specific certificate coverage, helpfulness, and false refusal. Automatic judging was validated on 600 human-annotated responses, achieving harmful-compliance F1 of 0.92, refusal accuracy of 0.93, helpfulness agreement of 0.87, and human Krippendorff's $\alpha = 0.80$. Results use three training seeds, $10{,}000$ base-request bootstrap samples, paired randomization tests, and Holm correction. Complete settings are in Appendix D.

6 Results

6.1 Singleton Robustness Does Not Compose

Across the three primary model families, $\mathrm{SCFR}_{6:8}$ ranges from 18.5% for Llama-3.1-8B-Instruct to 25.0% for Gemma-3-12B-IT, averaging 21.7%. Thus, more than one fifth of harmful requests that remain safe under the base prompt and all eight singleton factors fail under at least one strictly test-only composition. The phenomenon persists at larger scale: Qwen3-32B obtains $\mathrm{SCFR}_{6:8} = 15.0\%$ and $\mathrm{ASR}_{6:8} = 34.0\%$.

Maximum lower-order adverse curvature predicts test-only failure with AUROC 0.85 ± 0.02 and average precision 0.63, compared with $0.67 \pm 0.03/0.40$ for the minimum singleton margin and $0.70 \pm 0.02/0.44$ for cumulative singleton degradation. On Qwen3-32B, curvature remains predictive with AUROC 0.81, indicating that the effect is not confined to the three smaller models.

Which factors interact adversely? Figure 2 aggregates pair-specific adverse curvature over singleton-safe test prompts. The strongest interactions occur between claimed authority and benign-purpose rationalization (0.083), benign-purpose rationalization and structured decomposition (0.077), structured decomposition and output formatting (0.073), and role-play and structured decomposition (0.071). Thus, compositional vulnerability is not attributable to one universally strong transformation:

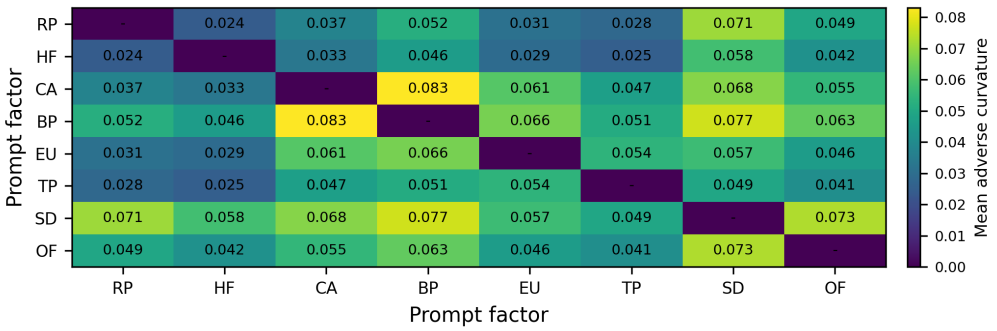

Figure 2: **Pair-specific adverse safety curvature.** Entries show mean $\epsilon_{ij}^{\theta}(x)$ over singleton-safe test prompts, averaged across the three primary model families. Larger values indicate stronger interaction-driven margin degradation. RP: role-play; HF: hypothetical framing; CA: claimed authority; BP: benign-purpose rationalization; EU: emotional urgency; TP: third-person framing; SD: structured decomposition; OF: output-format constraint.

Table 1: **Strict higher-order compositional robustness averaged over three model families.** $\text{SCFR}_{6:8}$ and $\text{ASR}_{6:8}$ use only test-only orders six through eight. Cert@3 uses the pair-specific certificate.

Method	$\text{SCFR}_{6:8}\downarrow$	$\text{ASR}_{6:8}\downarrow$	Cert@3 $\uparrow$	Help. $\uparrow$	Ref. $\downarrow$
Base model	21.7	40.6	24.0	80.4	6.1
Single-factor tuning	19.0	38.0	29.5	80.2	6.4
Pairwise training	16.0	34.0	37.5	80.0	6.6
Random composition	14.0	31.0	45.5	79.7	7.0
Hard-negative training	13.0	29.0	49.5	79.6	7.2
ASCoT-inspired	12.0	27.5	52.5	79.7	7.0
Margin hinge	13.0	28.5	50.5	79.9	6.7
Active margin projection	11.0	25.5	58.5	80.0	6.6
Curvature hinge	9.5	22.0	65.5	79.9	6.7
C^2SA	**7.5**	**18.0**	**75.5**	**80.1**	**6.3**

rationalization and decomposition instead act as interaction hubs whose effects depend on the other factors already present.

6.2 C^2SA Improves Strict Higher-Order Robustness

Table 1 compares methods under matched order support and editable-parameter budgets.

Relative to curvature hinge, C^2SA reduces $\text{ASR}_{6:8}$ by 4.0 points (18.2% relative; paired 95% CI $[-4.8, -3.2]$) and $\text{SCFR}_{6:8}$ by 2.0 points (21.1% relative; CI $[-2.6, -1.4]$). Relative to active margin projection, which uses the same active-set QP without curvature constraints, the reductions are 7.5 ASR points and 3.5 SCFR points. These controls indicate that the gains are not explained solely by higher-order data exposure, active constraint selection, or minimum-change optimization. Helpfulness changes from 80.4 to 80.1, while false refusal changes from 6.1% to 6.3%.

6.3 Mechanism Ablations and Margin Validity

Restricting curvature to $S = \varnothing$ misses interactions that emerge after other factors accumulate. Random constraint selection increases ASR from 18.0% to 22.4%, while replacing the minimum-change projection with a penalty update degrades both robustness and utility. Removing the utility constraint slightly lowers ASR but decreases helpfulness by 4.1 points and nearly doubles false refusal. Increasing the maximum curvature-context size from zero to three decreases $\text{ASR}_{6:8}$ from 29.2% to 18.0%; the intermediate values are 25.4% and 21.5% for context sizes one and two.

Table 2: **Mechanism ablations.** Lower SCFR, ASR, and false refusal are better; higher helpfulness is better.

Variant	$\text{SCFR}_{6:8} \downarrow$	$\text{ASR}_{6:8} \downarrow$	Help. ↑	Ref. ↓
Singleton constraints only	13.6	29.2	80.3	6.3
Curvature only at $S = \varnothing$	11.5	25.8	80.2	6.3
Random constraint selection	9.7	22.4	80.0	6.5
Penalty update without projection	8.8	20.6	79.1	7.6
No benign-utility constraint	**7.2**	**17.4**	76.0	12.4
Full C^{2}SA	7.5	18.0	**80.1**	**6.3**

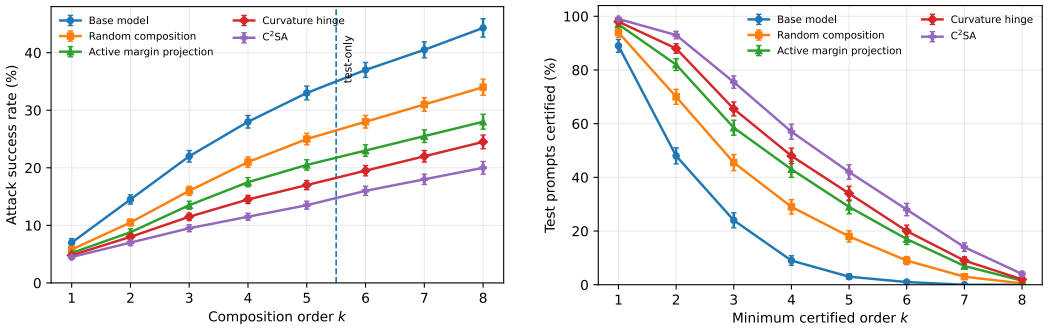

Figure 3: **Strict higher-order robustness and pair-specific certification. Left:** ASR by composition order; nodes through order five are within training support, whereas orders six through eight are excluded from training and model selection. **Right:** pair-specific certificate coverage after the final checkpoint is frozen. Curves show means over three model families and three seeds. Error bars denote 95% base-request bootstrap intervals.

Validity of the response margin. Across held-out nodes, the margin predicts policy-appropriate free generation with AUROC 0.88 ± 0.01 and Spearman correlation $\rho = 0.71$. Compositions with a positive certified lower bound produce appropriate responses in 93.4% of repeated generations, compared with 64.8% when the lower bound is negative. A negative bound means uncertified rather than necessarily unsafe.

6.4 Higher-Order Transfer and Certification

Figure 3(a) shows ASR by composition order. At orders six, seven, and eight, C^{2}SA obtains 16.0%, 18.0%, and 20.0%, respectively, compared with 19.5%, 22.0%, and 24.5% for curvature hinge and 23.0%, 25.5%, and 28.0% for active margin projection. The gap therefore persists throughout the test-only region.

Pair-specific certification is tighter than the uniform bound. For C^{2}SA, Cert@3 and Cert@4 are 75.5% and 57.0%, compared with 66.0% and 47.0% under the uniform scalar. Median certified order increases from three to four. The median observed-margin minus certified-bound gap is 0.14 (interquartile range $[0.08, 0.23]$), compared with 0.26 ($[0.15, 0.39]$) for the uniform certificate.

6.5 Out-of-Lattice and Compute-Matched Robustness

Table 3 evaluates robustness beyond the certified canonical lattice. These tests change factor realizations, factor families, rendering order, harm categories, or adaptively search against the repaired model; none is covered by the finite-lattice certificate.

Across the four distribution shifts, C^{2}SA lowers ASR by 4.0–5.0 points relative to curvature hinge and by 5.5 points under the adaptive attack. The gain is also consistent across model families: relative to curvature hinge, C^{2}SA reduces $\text{SCFR}_{6:8}/\text{ASR}_{6:8}$ from 7.0/18.5 to 5.0/14.5 on Llama, 9.5/22.0

Table 3: **Out-of-lattice and adaptive robustness (ASR, %).** All evaluations are empirical. Lower is better.

Method	Unseen tmpl.	Held-out factors	Order perm.	Unseen harms	Adaptive
Random composition	31.5	34.0	32.5	31.0	36.0
Active margin projection	28.0	31.0	29.5	28.0	30.5
Curvature hinge	25.5	28.5	26.5	25.5	27.5
C^2SA	**21.0**	**23.5**	**22.0**	**21.5**	**22.0**

Table 4: **Compute-matched strict higher-order robustness.** Total forward-plus-backward FLOPs are matched within 3% of C^2SA.

Method	Relative FLOPs	$SCFR_{6:8} \downarrow$	$ASR_{6:8} \downarrow$
Random composition	1.00	14.2	31.2
Hard-negative training	1.01	13.1	29.2
ASCoT-inspired	1.02	12.2	27.7
Active margin projection	1.01	11.1	25.7
Curvature hinge	1.02	9.6	22.2
C^2SA	1.00	**7.6**	**18.2**

to $7.5/18.0$ on Qwen, and $12.0/25.5$ to $10.0/21.5$ on Gemma. Paired base-request bootstrap intervals with Holm correction are reported in Appendix D.3.

Because constraint discovery adds forward scoring, we additionally match total forward-plus-backward computation within 3%. Table 4 shows that the method ordering is unchanged.

Under matched computation, C^2SA retains a 4.0-point ASR advantage and a 2.0-point SCFR advantage over curvature hinge. On independently generated H4RM3L- and CHIMERA-style attacks, aggregate ASR decreases from 40.0% for the base models to 22.0% for curvature hinge and 18.5% for C^2SA.

7 Limitations and Conclusion

Limitations. Our certificate applies to an appropriate-response margin over a declared finite lattice of policy-preserving transformations. It does not certify arbitrary jailbreaks, unverified factor families, alternative application orders, or every stochastic completion. The guarantee assumes that composed prompts preserve the requested action and response policy; manual validation reduces but cannot eliminate semantic ambiguity. Exact verification scales exponentially with the number of factors, restricting certification to moderate K. We audit the failure mode and curvature signal at 32B scale but do not run the full alignment procedure at that scale; scaling repeated constraint search and projection to substantially larger backbones remains future work. Finally, C^2SA uses local trust-region updates and response-anchor margins, so it provides neither a global optimization guarantee nor protection against deficiencies in the anchor construction.

Conclusion. We identify a singleton-safe, composition-unsafe failure mode in which individually tolerated prompt transformations interact to cross an aligned model's safety boundary. Adverse safety curvature measures these interactions, and factor-pair-specific bounds yield tighter lower bounds and certified orders than a uniform worst-case curvature scalar. Guided by this structure, C^2SA improves robustness on strictly unseen composition orders, unseen factor families, and adaptive attacks while preserving benign utility. These results expose a limitation of transformation-wise safety evaluation: *safe components need not compose safely*.

Reproducibility Statement

Appendix A describes the base-request splits, transformation factors, canonical compositor, admissibility criteria, response anchors, and human validation protocol. Appendix B provides the formal definitions and complete proofs of the pair-specific compositional margin bound and certificate.

Appendix C specifies the constraint-generation procedure, slack-regularized projection, editable parameterization, model-change metric, and optimization algorithm, while Appendix D reports all model, training, decoding, judging, statistical, and compute settings. Appendix E provides per-model results, context-depth and mechanism ablations, certificate-tightness analyses, compute-matched comparisons, and out-of-distribution and adaptive evaluations. The supplementary material includes the exact data splits, factor templates and held-out realizations, canonical compositor, admissibility annotations, response anchors, judge prompts, model revisions, random seeds, LoRA and trust-region configurations, and scripts used to generate every table and figure. We additionally report node and gradient evaluations, forward and backward FLOPs, wall-clock time, residual constraint violations, and slack statistics for all trainable methods.

USE OF LARGE LANGUAGE MODELS

Generative AI tools were used for prose editing, and manuscript organization. They were not used to execute experiments, access raw model outputs, or generate, modify, estimate, or interpolate any reported numerical result. The authors independently verified the mathematical derivations, implementation, citations, experiments, statistical analyses, figures, and claims, and take full responsibility for the entire content of the paper.

ETHICS STATEMENT

This work studies harmful requests and compositional jailbreaks in order to improve language-model safety. A C^2SA certificate concerns only the anchor-based appropriate-response margin over a specified, exhaustively verified finite transformation lattice. It does not certify arbitrary jailbreaks, every stochastic completion, semantic completeness, or deployment safety, and should not be presented to downstream users as a universal safety guarantee. Releasing the complete attack corpus could also facilitate misuse. Human annotators are informed in advance that the study contains potentially harmful material, may skip examples without penalty, and are provided with appropriate support and debriefing procedures.

REFERENCES

Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2–3):145–373, 2013.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Reinelle Jan Bugnot, Soohyeon Choi, Hoon Wei Lim, and Yue Duan. Compositional jailbreaking: An empirical analysis of mutator chain interactions in aligned LLMs. *arXiv preprint arXiv:2605.15598*, 2026.

Landon Butler, Abhineet Agarwal, Justin Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

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, 2024.

Mahavir Dabas, Tran Huynh, Nikhil Billa, Jiachen Tianhao Wang, Peng Gao, Charith Peris, Yao Ma, Rahul Gupta, Ming Jin, Prateek Mittal, and Ruoxi Jia. Adversarial déjà vu: Jailbreak dictionary learning for stronger generalization to unseen attacks. In *International Conference on Learning Representations*, 2026.

Moussa Koulako Bala Doumbouya, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky, and Christopher Manning. h4rm3l: A language for composable jailbreak attack synthesis. In *International Conference on Learning Representations*, 2025.

Satoru Fujishige. *Submodular Functions and Optimization*, volume 58 of *Annals of Discrete Mathematics*. Elsevier, 2 edition, 2005.

Jiabao Ji, Bairu Hou, Alexander Robey, George J. Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. Defending large language models against jailbreak attacks via semantic smoothing. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 7–40, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and the Association for Computational Linguistics.

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Bin Yu, and Kannan Ramchandran. SPEX: Scaling feature interaction explanations for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28878–28903. PMLR, 2025.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying LLM safety against adversarial prompting. In *Conference on Language Modeling*, 2024.

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.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Leo Hyun Park, Juwon Cho, Gyuhwan Kim, YoonDong Yeo, and Taekyoung Kwon. Chimera: Compositional jailbreak attacks on LLMs via judgment-driven search over heterogeneous strategies. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33330–33355, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *Transactions on Machine Learning Research*, 2025.

Zhengyue Zhao, Yingzi Ma, Somesh Jha, Marco Pavone, Patrick McDaniel, and Chaowei Xiao. ARMOR: Aligning secure and safe large language models via meticulous reasoning. In *International Conference on Learning Representations*, 2026.

Zihao Zhu, Xinyu Wu, Gehan Hu, Siwei Lyu, Ke Xu, and Baoyuan Wu. AdvChain: Adversarial chain-of-thought tuning for robust safety alignment of large reasoning models. In *International Conference on Learning Representations*, 2026.

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.