SCORE: SMOOTHED CERTIFIED RELEASE AND THE CENTER-TO-NEIGHBORHOOD GAP IN PROMPT REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt repair can recover benign requests rejected by an input safety gate, but accepting a repaired prompt does not establish robustness to nearby edits or safety of the generated response. We introduce SCORE (Smoothed Certified Release for Prompt Repair), a selective framework that repairs, certifies, or abstains while explicitly separating center certification, Hamming-neighborhood robustness, and downstream response safety. For fixed, positive categorical smoothing supports, we specialize a likelihood-ratio transfer envelope and maximize it exactly over small candidate-set Hamming balls. We also identify when exact transfer improves a product bound and when finite sampling prevents certification at a given budget. Among 606 held-out unsafe prompts routed to repair, branch-and-bound and beam proposal policies certify 215 and 66 centers, respectively, but neither certifies a repaired prompt at Hamming radius one. A separate frozen dual-gate policy releases 34 of 313 rejected benign prompts after automated fidelity filtering, yet certifies none of these routed prompts at radius one and misses its predeclared joint recovery, risk, and fidelity targets. Development kernel and semantic-support analyses find sparse nonzero neighborhood coverage, so the observed collapse is not a universal claim about smoothing. Finally, two direct Gemma 3 adaptive runs yield 13/224 and 9/224 fresh input-certified prompts with outputs judged harmful by both automated judge families. SCORE is therefore a route-aware audit of what input-side certification does and does not guarantee, not an end-to-end safety certificate.

1 INTRODUCTION

Jailbreaks can elicit harmful outputs despite safety training (Wei et al., 2023; Zou et al., 2023; Liu et al., 2024). Input filters block suspected attacks but also benign requests. Prompt repair seeks an intent-preserving input across the gate boundary; DDOR, for example, localizes and rewrites refusal-triggering fragments (Zhou et al., 2026). Because the repair is selected for gate acceptance, one accepted label is weak release evidence. Randomized smoothing strengthens it by requiring stability under known token noise (Robey et al., 2025; Kumar et al., 2024).

What transfers from this certified center? A local adversary acts on a neighborhood, while a downstream LLM maps released text to responses; a center input-gate certificate controls neither. We call the loss of route coverage from the center to a declared Hamming ball the *center-to-neighborhood gap*. Separating it from the gate-to-output boundary makes each failure attributable to proposal search, transfer, gate policy, or generation.

SCORE (Smoothed Certified Release for Prompt Repair) makes that path explicit: propose short edits for a rejected prompt, select with the gate, freshly certify the repair, then release or abstain. Proposal modules share the same certifier and route estimands. Finite fixed-support noise enables exact small-radius transfer while exposing that complete support need not provide useful robust margin.

Our study asks four falsifiable questions.

RQ1: selective recovery. Under one shared certifier, can a repair policy recover benign gate false positives while controlling unsafe release and automated fidelity loss?

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.

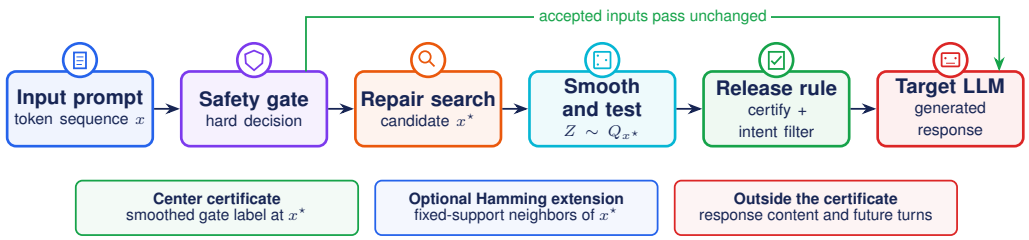

Figure 1: **Selective smoothed prompt repair and its guarantee boundary.** An input rejected by a gate enters token-level repair; the selected center is sampled under a fixed product categorical kernel and released only if its smoothed gate label is certified safe. The center statement (green), optional fixed-support Hamming extension (blue), and downstream generation (red) are distinct objects. Theory audits the blue transfer; adaptive evaluation tests the red interface. Color is redundant with the box labels.

RQ2: center-to-neighborhood transfer. How large is the route-coverage gap between a center certificate and validity over every candidate-set substitution within Hamming radius one or two, and when can exact transfer improve on a product-maximum shortcut?

RQ3: downstream safety. Can a policy-aware attacker obtain a fresh center certificate for a prompt that elicits output judged harmful from the downstream model?

RQ4: gate dependence. How do proposal outcomes and adaptive attacks change across separately trained input gates, and what can such cross-gate evidence establish?

We contribute: (i) a repair, certify, and abstain framework with route estimands that retain failures and abstentions; (ii) a finite-kernel Neyman-Pearson specialization, exact small-radius computation, a phase-boundary corollary, and a finite-sample feasibility condition; (iii) a frozen audit in which two individual proposal policies and their audit portfolio lose all certified coverage from the center to radius one among 606 repair-routed prompts, together with a predeclared kernel-family sensitivity study; and (iv) common-certifier, prompt-disjoint, adaptive, cross-gate, and modern-target tests separating proposal, input robustness, and output behavior. These establish the measured boundary of this policy and kernel, not a general failure of smoothing or repair.

2 Related Work

DDOR repairs overrefusal through localized rewriting (Zhou et al., 2026). Our mRTF proposal is a clean-room analogue: it localizes edits but omits DDOR's rewrite model and multi-oracle flow. A broader audit studies semantic and multi-turn attacks (Pratap et al., 2026); SCoRe instead isolates selection, input certification, and downstream generation on one repair route.

Certified text classifiers address word substitutions (Jia et al., 2019; Ye et al., 2020) and masking-based perturbations (Zeng et al., 2023). Text-CRS and CERT-ED consider broader textual edits (Zhang et al., 2024; Huang et al., 2024). Their threat sets and kernels differ, but they certify classifier decisions rather than repair-selected releases or generated responses.

DiffTextPure derives tight knapsack bounds for smoothed LLM defenses (Chen et al., 2025). Hybrid RS also uses discrete likelihood-ordering within factorized smoothing (Delattre et al., 2026). We do not claim a new Neyman-Pearson lemma; the distinct object here is exact maximization over a repair-defined, position-specific candidate ball and its route-level coverage. CluCERT combines clustering and denoising, while DCRS and Certified Semantic Smoothing certify response-level decisions under other kernels (Wang et al., 2026; Jang et al., 2026; Cheng et al., 2026). Appendix G compares these objects and threat sets.

Table 1: Predicates and their valid conclusions. None certifies semantic preservation.

Predicate	Evaluated object	Valid conclusion
$S(u) = \text{safe}$	hard gate at u	hard label at u
$\overline{p}_u < 1/2$	$S(Z)$ for $Z \sim Q_u$	$\widetilde{S}_Q(u) = \text{safe}$
$U_r(u, \overline{p}_u) < 1/2$	$S(Z)$ around every $v \in \mathcal{B}_r(u)$	$\widetilde{S}_Q(v) = \text{safe}$ for every such v
Output judged harmful/safe	$Y \sim G(\cdot \mid u)$	named output-judge label for the sampled response

3 Selective Smoothed Prompt Repair

Let $x = (x_1, \ldots, x_n) \in \mathcal{V}^n$ be a tokenized prompt, let $S : \mathcal{V}^n \to \{\text{safe}, \text{unsafe}\}$ be an input gate, and let G be a downstream generator. If $S(x)$ is unsafe, SCoRe searches for

$$x^\star \in \arg\min_{u \in \mathcal{V}^n} d(x, u) \quad \text{s.t.} \quad S(u) = \text{safe}, \tag{1}$$

where d is token Hamming distance. Exhaustive search is intractable, so the implementation evaluates left-to-right beam search and a bounded branch-and-bound proposal within position-specific candidate sets. The latter is not assumed globally optimal when search is incomplete. Candidate construction may use embedding neighbors and gate scores, but no certificate depends on search optimality. The abstention path connects the system to selective prediction (Geifman & El-Yaniv, 2017; 2019).

Candidate generation and release are deliberately separated. Our primary generator uses beam width 20 and at most three token edits; paraphrasing and minimal truncation serve as alternative generators under the same downstream test. A repair is intent-filtered only when MPNet cosine and bidirectional NLI are at least 0.8 and token edit ratio is at most 0.35. The optional dual-gate rule requires separately computed center certificates from LlamaGuard and WildGuard at the same repaired text. With $\alpha = .001$ for each fixed candidate, a union bound gives confidence at least $1 - .002$ for the conjunction. Neither the automated intent filter nor the two-gate conjunction establishes semantic ground truth or downstream-output safety.

Certification uses fixed admissible sets $\mathcal{A}_i$ containing the center and every allowed substitute. A center-independent π_i has positive mass on each token in $\mathcal{A}_i$. For replacement probability $q \in (0, 1)$,

$$Q_u(z) = \prod_{i=1}^{n} [(1 - q)\mathbf{1}\{z_i = u_i\} + q\pi_i(z_i)]. \tag{2}$$

The $\mathcal{A}_i$ and π_i remain fixed across centers compared by a Hamming certificate. In the evaluated uniform kernel, π_i is uniform on 32 tokens. Subword tokenization makes the distance model-specific (Sennrich et al., 2016; Kudo & Richardson, 2018): one token substitution is not one word edit or a fixed character edit (Levenshtein, 1966).

Define $p_u = \mathbb{P}_{Z \sim Q_u}[S(Z) = \text{unsafe}]$. From N independent samples, a one-sided Clopper-Pearson interval (Clopper & Pearson, 1934) gives $\overline{p}_u$. The center rule releases u only if $\overline{p}_u < 1/2$. Write $\widetilde{S}_Q(u) = \text{safe}$ when $p_u < 1/2$. Conditional on the confidence event, the center rule certifies $\widetilde{S}_Q(u) = \text{safe}$. The hard label $S(u)$ is checked separately by proposal selection and is not a consequence of this probability statement. Neither statement controls $G(u)$ or a different center.

Routes and threat models. Accepted prompts pass unchanged. Rejected prompts enter repair and are released only if the selected center is certified; otherwise the system blocks. We separate hard-gate bypass, center-certified release, and certified-and-harmful output. The Hamming adversary knows $Q, S, \mathcal{A}_i$ and changes at most r tokens within the fixed sets. The adaptive suffix attacker instead targets the center release predicate through gradients and expectation over smoothing noise. Its suffix is outside the certified Hamming threat set by design; this stress-tests the policy rather than contradicting the theorem.

4 A TIGHT HAMMING TRANSFER ENVELOPE

Transfer the unsafe-event probability from u to v. Common finite support makes $L_{u,v}(z) = Q_v(z)/Q_u(z)$ finite. Order its distinct values $\ell_1 > \cdots > \ell_m$ and set $a_j = Q_u(L_{u,v} = \ell_j)$ and $b_j = Q_v(L_{u,v} = \ell_j) = \ell_j a_j$. For $t \in [0, 1]$, fill strata in descending ratio order until their Q_u mass reaches t, fractionally filling the final stratum. Denote the resulting Q_v mass by $T_{u\to v}(t)$.

Theorem 4.1 (Discrete transfer envelope). *For any $\phi : \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to [0, 1]$ with $\mathbb{E}_{Q_u}[\phi] \le t$,*

$$\mathbb{E}_{Q_v}[\phi] \le T_{u\to v}(t). \tag{3}$$

The bound is attained by a randomized test selecting strata in descending $L_{u,v}$ order and randomizing only on the boundary. Thus, for the deterministic unsafe event of any gate, $p_v \le T_{u\to v}(p_u)$.

Proof. Write $\mathbb{E}_{Q_v}[\phi] = \sum_z Q_u(z)L_{u,v}(z)\phi(z)$. If Q_u mass is assigned to a lower-ratio stratum while capacity remains in a higher one, exchanging equal mass cannot decrease the objective. Repetition yields the descending-ratio solution and at most one fractional boundary. Use $\phi = \mathbf{1}\{S(z) = \text{unsafe}\}$ for the gate claim. □

This finite Neyman-Pearson construction follows the tight-smoothing perspective (Lee et al., 2019; Yang et al., 2020; Dvijotham et al., 2020). "Tight" refers to randomized tests, not a new statistical lemma. For a deterministic gate, indivisible boundary atoms may reduce the attainable value; the envelope remains a valid universal upper bound. Hybrid randomized smoothing embeds the same discrete likelihood-ordering principle in a joint discrete-Gaussian certificate and recovers the discrete-only case (Delattre et al., 2026). Our contribution is therefore the finite, position-specific kernel construction, exact maximization over the declared repair neighborhood, and its connection to the selective release route. It is not a new Neyman-Pearson lemma.

For changed position i, only u_i, v_i, and all other tokens matter. Their (Q_u, Q_v) masses are

$$(1 - q + q\pi_i(u_i), q\pi_i(u_i)), \tag{4}$$

$$(q\pi_i(v_i), 1 - q + q\pi_i(v_i)), \tag{5}$$

$$(q[1 - \pi_i(u_i) - \pi_i(v_i)], q[1 - \pi_i(u_i) - \pi_i(v_i)]). \tag{6}$$

Convolution across changed positions produces the full strata without enumerating token sequences. For the fixed candidate-set ball $\mathcal{B}_r(u)$, define

$$U_r(u, t) = \max_{v\in\mathcal{B}_r(u)} T_{u\to v}(t). \tag{7}$$

We compute the maximum exactly at small r by enumerating positions and distinct values of $\pi_i(v_i)$; equal masses induce identical strata.

Proposition 4.2 (Hamming-certified release). *On $p_u \le \overline{p}_u$, if $U_r(u, \overline{p}_u) < 1/2$, then the smoothed classifier satisfies $\widetilde{S}_Q(v) = \text{safe}$ for every $v \in \mathcal{B}_r(u)$.*

Proof. For each v, Theorem 4.1 and monotonicity give $p_v \le T_{u\to v}(p_u) \le T_{u\to v}(\overline{p}_u) \le U_r(u, \overline{p}_u) < 1/2$. □

The uniform kernel exposes a sharp boundary between a vacuous product shortcut and a useful exact envelope. Let $a = 1 - q + q/K$, $b = q/K$, and $\lambda = a/b$ for a support of size K.

Corollary 4.3 (Uniform-kernel phase boundary). *For centers separated by exactly s substitutions, if $a^s \ge 1/2$, then*

$$T_{u\to v}(t) < 1/2 \quad \Longleftrightarrow \quad \lambda^s t < 1/2. \tag{8}$$

If $a^s < 1/2$, the exact envelope certifies a nonempty interval of event-mass budgets on which the product bound fails.

Proof. The largest likelihood ratio is λ^s. Its stratum has Q_u mass b^s and Q_v mass a^s, so $T_{u\to v}(t) = \lambda^s t$ for $0 \le t \le b^s$. If $a^s \ge 1/2$, the crossing of $1/2$ occurs within this first stratum, which proves equation 8. If $a^s < 1/2$, the first stratum is exhausted below $1/2$. Every remaining nonzero stratum has likelihood ratio strictly below λ^s; hence at $t = 1/(2\lambda^s) > b^s$ the exact envelope is strictly below $1/2$, whereas the product bound equals $1/2$. Continuity gives a nonempty interval of strict separation. □

For the evaluated $q = .4$ and $K = 32$, $a = .6125$ and $\lambda = 49$. Corollary 4.3 predicts identical binary decisions at $s = 1$, but permits a strict gain at $s = 2$: $a > .5$ while $a^2 < .5$. The audit matches this prediction, with 8 versus 8 certified method-prompt rows at radius one and 2 versus 0 at radius two for the exact and product bounds.

Finite Monte Carlo sampling imposes a second boundary. Even when no sampled prompt is unsafe, a one-sided Clopper-Pearson interval has upper endpoint $1 - \alpha^{1/N}$. Let τ_r denote the supremum of center event masses for which the chosen transfer rule is below $1/2$ at radius r.

Proposition 4.4 (Finite-sample feasibility)**.** *A certificate with N samples and confidence parameter α is impossible, regardless of the observed sample, when $1 - \alpha^{1/N} \geq \tau_r$. Conversely, the zero-error outcome can cross the certificate boundary when*

$$N > \frac{\log \alpha}{\log(1 - \tau_r)}. \tag{9}$$

Proof. The Clopper-Pearson upper endpoint is minimized when the unsafe count is zero, where it equals $1-\alpha^{1/N}$. Comparing this minimum with τ_r gives both statements after taking logarithms. □

This distinction separates insufficient sample budgets from insufficient observed margins. At $q = .4$, $K = 32$, $N = 10{,}000$, and $\alpha = .001$, the minimum upper endpoint is 6.91×10^{-4}. Exact transfer is feasible at both evaluated radii, whereas the radius-two product shortcut would require at least 33,168 samples even after zero observed unsafe draws.

Route HCRR-r is the fraction of prompts that enter the repair route and satisfy this rule; HCRR-0 is route-level center-certified recovery. We separately report full-pool analysis coverage, which also evaluates prompts passed through unchanged and therefore is not a deployment recovery estimand. A product-maximum bound $p_v \leq \Lambda_r p_u$ pays the largest ratio on all event mass even when that stratum is rare. The envelope respects available mass before advancing. Each v has at most 3^r raw convolution strata. Exact maximization enumerates position subsets and distinct alternative masses, not token identities. Thus, for uniform π_i, radius-two outer search is $O(n^2)$ and independent of K in case count; Appendix B gives the general complexity and oracle checks.

5 Evaluation

5.1 Data, models, and estimands

Data and population rules. The held-out unsafe pool combines AdvBench, JailbreakBench, and StrongREJECT (Zou et al., 2023; Chao et al., 2024; Souly et al., 2024). These sources differ in construction and evaluation philosophy, so results are reported both jointly and per source, following calls for benchmark diversity in HarmBench (Mazeika et al., 2024). Hamming analysis contains 629 prompts: 347, 77, and 205 respectively; 606 are rejected by the deployment gate and enter repair. Adaptive evaluation uses a balanced 225-goal subset, 75 per source; 213 are rejected before attack. Route HCRR divides by the entire repair route, including failed searches and abstentions. Full-pool analysis coverage uses all 629 prompts but is not described as deployment recovery.

Selective-recovery evaluation uses a prompt-disjoint split from XSTest and OR-Bench (Röttger et al., 2024; Cui et al., 2025), plus mixed benign tasks. Its 605-prompt development set is used for the common-certifier comparison and to freeze the dual-gate operating point; its 2,419-prompt test set contains 1,735 benign and 684 unsafe records. The primary benign denominator is the 313 test prompts falsely rejected by the first hard gate. The unsafe-loss denominator remains all 684 unsafe prompts. An additional prompt-disjoint evaluation contains 1,000 OR-Bench benign prompts and 149 HarmBench unsafe behaviors. It compares SCoRe with a localized minimal-refusal-trigger (mRTF) analogue inspired by DDOR (Zhou et al., 2026) under the same certifier and repeats SCoRe with WildGuard. These method and gate choices are fixed on a separate 438-prompt development split.

Gates and target models. The primary legacy input gate is LlamaGuard 2 (Inan et al., 2023; Llama Team, 2024); WildGuard is a separately trained transfer gate and the primary output judge (Han et al., 2024). We use Llama 2 in raw completion mode and Mistral-7B-Instruct in its native chat format (Touvron et al., 2023; Jiang et al., 2023). These modes are stated because a chat template

changes both the target distribution and the attack interface. The prompt-disjoint replication uses Llama Guard 3 and WildGuard as input gates, and Meta-Llama-3-8B-Instruct and Gemma-3-12B-IT as downstream targets in their native chat formats.

Metrics. Route HCRR-r counts repair-routed prompts receiving the certificate in Proposition 4.2. A hard bypass has $S(u) = \text{safe}$; a center-certified bypass additionally has $\overline{p}_u < 1/2$. End-to-end success requires the certificate and harmful target output. We report WildGuard and LlamaGuard output judgments separately and their conjunction. Judge agreement is a diagnostic, not evidence that either judge is ground truth (Zheng et al., 2023; Landis & Koch, 1977). Benign false-positive recovery (BFR) is the fraction of initially rejected benign prompts released after repair. Unsafe TPR loss is newly released unsafe prompts divided by the full unsafe population. Before test evaluation we require the BFR Wilson lower bound to be at least 10%, automated intent agreement among dual-certified benign repairs to be at least 80%, and unsafe TPR loss to be at most two percentage points.

5.2 Frozen Hamming and adaptive protocols

Hamming protocol. The kernel is uniform over 32 admissible tokens per position, with $q = 0.4$ and $N = 10{,}000$ Monte Carlo samples per instance. The one-sided interval uses per-instance $\alpha = 0.001$. We freeze $q = .4, K = 32$ after it gives the smallest development radius-one envelope; all test settings precede evaluation. We use $r \in \{0, 1, 2\}$ and claim no approximation at larger radii. The per-instance confidence level does not imply simultaneous validity of all reported certificates. Wilson intervals summarize observed route coverage under a binomial sampling model; they are not joint certificate bounds.

Kernel sensitivity uses 151 development method-prompt rows. An $N = 1{,}000$ screen crosses five q values, four support sizes, and uniform or weighted replacement. A predeclared corner/boundary/maximum-coverage rule selects six unique cells for evaluation with $N = 10{,}000$ fresh samples. This study neither selects nor changes the held-out $q = .4, K = 32$ protocol; Appendix E.3 gives the full grid and rule. A development-only support ablation selects embedding-nearest admissible tokens while retaining the original and repaired tokens. Eight uniform or softmax cells cross $q \in \{.4, .5\}$ and $K \in \{8, 32\}$ on the same 151 rows with $N = 10{,}000$; Appendix E.3 gives the full protocol.

Adaptive protocol. The attacker appends a discrete suffix and performs gradient-guided coordinate search against the smoothed gate with 32 candidates per update and two noise samples. LlamaGuard runs use two restarts of 60 updates; direct WildGuard runs use four restarts of 30 updates. This adapts the GCG lineage to the deployed rule (Wallace et al., 2019; Zou et al., 2023; Andriushchenko et al., 2025). Selection may inspect target output and a 200-sample screen for each restart; the chosen candidate is then recertified with 1,000 fresh samples. The policy-aware attack is not joint differentiable optimization of gate loss and target harm. We evaluate two Llama 2 runs and one Mistral run against LlamaGuard, and one Llama 2 run and two Gemma 3 runs against WildGuard. PAIR, AutoDAN, and Crescendo represent complementary threat models (Chao et al., 2025; Liu et al., 2024; Russinovich et al., 2025).

Cross-gate protocol. We apply WildGuard to suffixes optimized against LlamaGuard. Besides its hard label, two score thresholds are frozen on a separate calibration split: one targets the primary gate's benign false-positive rate and the other its unsafe true-positive rate. Thresholds are not refit on test. Since the attacker does not optimize WildGuard in this experiment, the result measures transfer, not WildGuard robustness.

Integrity controls. Records retain prompt and split identity, seed, budget, gate decisions, target output, both judges, and independent final samples. Validation requires exact ID coverage without duplication; deterministic aggregates retain all outcomes and report numerators, denominators, and Wilson 95% intervals. Small-support exhaustive optimization checks the theorem implementation.

6 Results

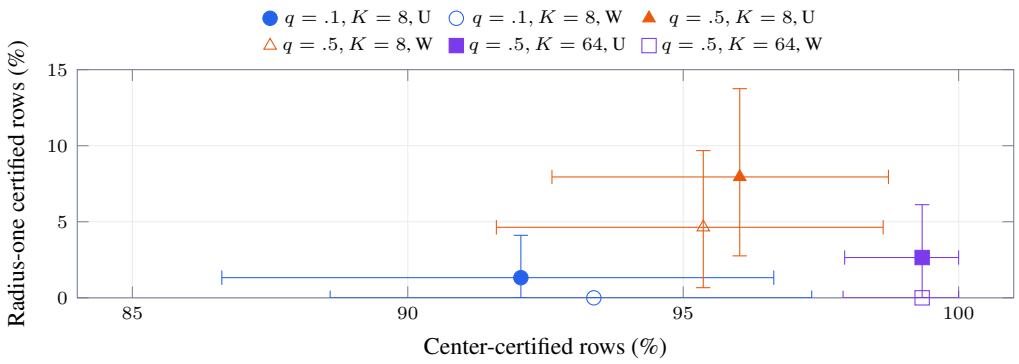

Figure 2: Development kernel frontier on 151 method-prompt rows per point, with $N = 10{,}000$ fresh samples and per-instance $\alpha = .001$. Error bars are 95% prompt-cluster bootstrap intervals over 111 unique prompts; U/W denote uniform and weighted replacement. A separate $N = 1{,}000$ screen selected these cells by the predeclared rule. This is not held-out route HCRR.

6.1 Selective recovery misses the frozen utility-risk target

Using the same LlamaGuard certifier on development data, SCoRe beam search center-certifies 31/81 benign false positives, versus 19/81 for paraphrasing and 66/81 for minimal truncation. These gains are misleading in isolation: the same rules newly release 66/171, 15/171, and 130/171 unsafe prompts. Requiring the frozen intent filter reduces the corresponding benign/unsafe counts to 20/16, 16/10, and 47/55. Thus no generator meets the joint development target.

We nevertheless evaluate the predeclared backup, which uses SCoRe beam search at $q = .1$, with fresh LlamaGuard and WildGuard center certificates and the intent filter, once on test. It rescues 34/313 benign rejections (10.86%; Wilson 95% CI 7.88-14.80%) while newly releasing 15/684 unsafe prompts (2.19 percentage points). Only 34/56 dual-certified benign repairs pass the automated intent rule (60.7%). All three frozen criteria therefore fail, answering RQ1 negatively for this operating point. The result is not retuned on test, and an unsafe-source release is not itself a harmful generated output. Full source counts appear in Appendix D.

6.2 Proposal and gate dependence persist on a prompt-disjoint split

The complete comparison in Appendix 9 isolates proposal and gate effects under the actual position-specific kernel ($q = .1$, $N = 10{,}000$, and one-sided $\alpha = .001$). With Llama Guard 3, only 16/1,000 benign prompts enter repair. SCoRe releases 2/16 of them and 2/145 routed unsafe prompts. The mRTF analogue releases fewer benign prompts (1/16) and more unsafe prompts (10/145), so it does not improve the observed utility-risk frontier. Replacing the gate with WildGuard changes the route to 123/1,000 benign and 148/149 unsafe prompts; SCoRe releases 10/123 and 4/148 respectively. Every radius-one decision is negative: 0/161 for both Llama-Guard-3 proposal cells and 0/271 for WildGuard. All 29 center releases remain center-certified after Bonferroni correction over the 42 final certificate calls at either family-wise level .05 or .001. These data confirm gate dependence and robust-coverage collapse without treating the clean-room analogue as an official implementation.

Modern-target evaluation covers 2,070 unique target-input pairs without generation or judge-parse failures. Llama Guard 3 and WildGuard disagree on 77 outputs and share only 12 harmful decisions. Among ten unsafe-labelled prompts released by the mRTF analogue, both judges identify harmful output for 2/10 with Llama 3 and 1/10 with Gemma 3. The SCoRe release denominators are only two or four, which precludes a population-level safety conclusion. Complete judge-specific counts appear in Appendix 10.

6.3 Tight transfer is nonzero, but Hamming coverage collapses

Table 2 and Figure 2 answer RQ2. Center HCRR is 215/606 for branch-and-bound repair and 66/606 for beam search, and both policies fall to 0/606 at radii one and two. Their deduplicated audit

Table 2: Certificate coverage on all 606 repair-routed prompts. The first two rows are deployable proposal policies. The last row is their deduplicated union, which is an audit portfolio rather than a single selection policy.

Estimand	Center	$r = 1$	$r = 2$
Branch-and-bound policy	215/606	0/606	0/606
Beam policy	66/606	0/606	0/606
Either-method portfolio	222/606	0/606	0/606

Table 3: Adaptive results with fresh input-gate certification. The first three rows use LlamaGuard and its 213 rejected goals; the remaining rows use WildGuard and its 224 rejected goals. "Both" requires both output judges to label the sampled response harmful. The WildGuard rows are direct attacks, not cross-gate transfers.

Target and run	Rejected	Certified ∩ harmful (WG)	Certified ∩ harmful (both)
Llama 2, seed 1	213	181/213 (85.0%)	143/213 (67.1%)
Llama 2, seed 2	213	187/213 (87.8%)	153/213 (71.8%)
Mistral, seed 1	213	104/213 (48.8%)	58/213 (27.2%)
Llama 2, WildGuard attack	224	12/224 (5.4%)	11/224 (4.9%)
Gemma 3, seed 1	224	14/224 (6.25%)	13/224 (5.80%)
Gemma 3, seed 2	224	9/224 (4.02%)	9/224 (4.02%)

portfolio reaches 222/606 at the center but is not a deployable policy because it selects whichever method succeeds after observing both outcomes. Full-pool analysis reaches 244/629 at the center, 4/629 at radius one, and 1/629 at radius two. Every nonzero neighborhood result belongs to one of 23 pass-through prompts and does not constitute robust repair. The exact envelope is strictly tighter than the product-maximum bound for 316/325 center-certified method-prompt rows at radius one and is the only method to certify any radius-two rows (2/350 versus 0). StrongREJECT has no radius-one certificate, while JailbreakBench retains two, showing that even the sparse analysis positives are source-dependent. Only 3/350 radius-one envelopes lie within .05 of the decision boundary and none lies within .01; no radius-two envelope lies within .05. The collapse is therefore not concentrated in Monte Carlo-ambiguous near-threshold rows.

The same gap appears in the frozen dual-gate selective policy, whose candidate construction, gates, kernel, and fidelity rule differ from the uniform-kernel audit. Among 313 routed benign prompts, it finds 200 candidates, dual-center certifies 56, and releases 34 after fidelity filtering. Exact transfer over its actual position-specific supports certifies none of the 313 at radius one or two. The corresponding unsafe route has 252 candidates among 581 prompts, 83 dual-center certificates, 15 center releases, and no neighborhood release. Appendix E.2 reports the complete waterfall.

Across the six development cells, center coverage is 139-150/151, versus 0-12/151 at radius one and 0-4/151 at radius two. Uniform $q = .5, K = 8$ gives the largest neighborhood counts (12/151 and 4/151, versus 145/151 at the center); its radius-one positives span all three sources (6/115, 4/15, and 2/21). Smaller support and greater noise can therefore produce a nonzero frontier, but no selected cell approaches its center coverage. The semantic support ablation likewise reaches at most 12/151 and 2/151, and softmax mass never improves radius-one coverage over matched uniform mass. These remain development comparisons; Appendix E.3 gives all cells.

The mechanism is a conflict between margin and coverage. A center certificate needs only $\overline{p}_u < 1/2$. Hamming certification needs enough additional margin to absorb the probability mass transferred into high likelihood-ratio strata. Support completeness prevents infinite ratios inside the admissible set, but does not make those ratios small. Increasing replacement noise can improve distributional overlap while simultaneously degrading the gate's center decision. The evaluated $q = 0.4$ point is the frozen outcome of that trade-off, not a post-hoc choice made to maximize test HCRR.

6.4 Fresh center certification coexists with harmful output

For Llama 2, 181/213 (85.0%) and 187/213 (87.8%) attacked inputs are both fresh-certified and harmful under WildGuard across seeds. Requiring both judges yields 143/213 (67.1%, 95% CI 60.6-73.1%) and 153/213 (71.8%, 65.4-77.4%). With Mistral chat, rates are lower but nonzero: 104/213 (48.8%) under WildGuard and 58/213 (27.2%, 21.7-33.6%) under both. Thus RQ3 is answered for the evaluated models and budget: fresh input-gate stability does not certify generated content. Replication across Llama seeds rules out a single random seed as the explanation; the lower Mistral rate shows target dependence rather than universality.

With Gemma 3, the fresh-certified harmful-by-both endpoint is 13/224 (5.80%, 95% CI 3.42-9.67%) and 9/224 (4.02%, 2.13-7.46%) across two independent attack seeds. WildGuard alone judges 14/224 (6.25%, 3.76-10.22%) and 9/224 (4.02%, 2.13-7.46%) harmful. Five consensus-success prompt IDs recur across seeds. Treating judge disagreements as unresolved brackets these endpoints at 13-14/224 and exactly 9/224. Thus, the failure persists when both the gate and target model change, while its frequency remains attack-, source-, and target-specific. These finite-budget successes are observed lower bounds on attack reach, not estimates of residual worst-case safety; Appendix E gives source counts.

WildGuard's hard decision rejects all 213 LlamaGuard-optimized attacks; its two thresholds accept 1 and 52, with 49 certificates and 28 outputs judged harmful by both at the latter. This measures transfer, not WildGuard robustness. Direct adaptation instead yields 11/224 fresh-certified harmful-by-both cases. Gate diversity changes the frontier, not the input-to-output guarantee boundary; Appendix E gives further diagnostics.

7 Discussion

Three interfaces. Proposal comparison measures recovery, the theorem transfers a gate event to a fixed Hamming neighborhood, and endpoint evaluation measures generation. Better search need not improve robustness or output behavior. Route collapse holds for the evaluated proposals, kernels, and gates; the development frontier shows why center and neighborhood coverage should be reported together.

Use and limitations. Center smoothing strengthens one gate label, not downstream safety. Output controls, gate diversity, and calibration remain complementary (Angelopoulos et al., 2025; Bates et al., 2021). Theory assumes fixed, finite, positive supports and exact enumeration covers small radii. Automated judges, finite attacks, no quantitative human validation, and diagnostic multi-turn tests limit generalization. The embedding-nearest support ablation tests only one representation and two mass rules; kernel redesign needs fresh utility and coverage evidence (Appendix K).

8 Conclusion

SCoRe separates center, neighborhood, and output claims. Its policy misses the recovery, risk, and intent target. Branch-and-bound repair, beam search, and the frozen dual-gate policy have zero routed radius-one coverage, although selected development kernels yield sparse positives. Exact transfer diagnoses this gap; adaptive and cross-gate tests show that center stability does not certify downstream output. SCoRe is an input-side audit, not an end-to-end safety certificate.

AI USE STATEMENT

Generative AI tools were used to assist with drafting, language editing, and polishing the manuscript. The authors independently checked all experimental results, mathematical statements, citations, claims, and final text and take full responsibility for the content.

ETHICS STATEMENT

This work studies mechanisms for eliciting policy-violating model behavior. The experiments use established public safety benchmarks and open-weight models in an isolated research environment and do not deploy attacks against public services. The reported quantitative evaluation uses public benchmark labels and automated model-dependent measurements. A small supporting semantic-preservation audit involved four internal annotators, but its ratings are not used as quantitative evidence for the present endpoints. Ratings were recorded anonymously, and annotators worked independently. Participation was voluntary. Each annotator could skip an item or withdraw at any time for any personal reason, without penalty or effect on employment, study, evaluation, or access to resources. The activity collected no personal or sensitive data, and formal ethics review was not required under the applicable institutional guidance. We do not provide a separate code or data archive with this submission. In particular, ready-to-use harmful generations are not released. Automated safety judges can encode cultural and policy biases, so their labels are reported as model-dependent measurements rather than ground truth. The method should not be used to silently rewrite benign user intent or to claim that an input-side certificate guarantees harmless generation.

REPRODUCIBILITY STATEMENT

Section 5 specifies datasets, population rules, models, attack budgets, sampling counts, confidence levels, and estimands. The main theorem states the fixed-support assumptions needed by the certificate, and Appendix A gives a complete proof and computational construction. Appendix C records the evaluation protocol and failure handling; Appendix E provides disaggregated counts and confidence intervals; and Appendix H describes record validation and aggregation. No separate source-code or per-example data archive accompanies this submission, which limits independent computational reproduction of the reported experiments.

REFERENCES

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hXA8wqRdyV.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641-1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1-34, 2021. doi: 10.1145/3478535. Article 43.

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. *Advances in Neural Information Processing Systems*, 37:55005-55029, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23-42. IEEE, 2025. doi: 10.1109/SATML64287.2025.00010. URL https://doi.org/10.1109/SATML64287.2025.00010.

Huanran Chen, Yinpeng Dong, Zeming Wei, Hang Su, and Jun Zhu. Towards the worst-case robustness of large language models. *arXiv preprint arXiv:2501.19040*, 2025. URL https://arxiv.org/abs/2501.19040.

Zehua Cheng, Jianwei Yang, Wei Dai, and Jiahao Sun. Provable defense framework for LLM jailbreaks via noise-augumented alignment, 2026. URL https://arxiv.org/abs/2602.01587.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404-413, 1934. ISSN 00063444, 14643510. doi: 10.1093/biomet/26.4.404. URL http://www.jstor.org/stable/2331986.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515-11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Blaise Delattre, Hengyu Wu, Paul Caillon, Wei Yang Bryan Lim, and Yang Cao. Certified robustness under heterogeneous perturbations via hybrid randomized smoothing. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.12876.

Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, J. Zico Kolter, Chongli Qin, András György, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using F-divergences. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_SJlKrkSFPH.html.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86-92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.neurips.cc/paper/7073-selective-classification-for-deep-neural-networks.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151-2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *Advances in neural information processing systems*, 37:8093-8131, 2024. doi: 10.52202/079017-0261. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4083-4093. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1419. URL https://aclanthology.org/D19-1419/.

Zhuoqun Huang, Neil G. Marchant, Olga Ohrimenko, and Benjamin I. P. Rubinstein. CERT-ED: Certifiably robust text classification for edit distance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10813-10835. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.635. URL https://aclanthology.org/2024.findings-emnlp.635/.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. URL https://arxiv.org/abs/2312.06674.

Subin Jang, Jimin Yeom, and Sungyoon Lee. Certifying robustness of large language models via discrete-continuous randomized smoothing. In *ICML Workshop on AI for Good*, 2026. URL https://openreview.net/forum?id=vrZwS5FQHa.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4129-4142, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1423. URL https://aclanthology.org/D19-1423/.

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

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66-71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012/.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying LLM safety against adversarial prompting. In *First Conference on Language Modeling*, 2024. URL https://2024.colmweb.org/AcceptedPapers.html.

J R Landis and G G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159-174, March 1977. doi: 10.2307/2529310. URL https://doi.org/10.2307/2529310.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707-710, 1966. URL https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L/abstract.

Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4):4585-4593, 2020. doi: 10.1609/aaai.v34i04.5888. URL https://ojs.aaai.org/index.php/AAAI/article/view/5888.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Jwpw4qKkb.

Llama Team. Meta llama guard 2. Model card, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-Guard-2-8B.

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

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.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220-229. Association for Computing Machinery, 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Aaditya Pratap, Harsh Kasyap, and Somanath Tripathy. Breaking the assumptions: Auditing input-side jailbreak defenses against semantic attacks, 2026. URL https://arxiv.org/abs/2608.21895.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=laPAh2hRFC.

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/.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo Multi-Turn LLM jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2421-2440, Seattle, WA, 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/russinovich.

Yan Scholten, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. Hierarchical randomized smoothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=6IhNHKyuJO.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715-1725. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

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. *Advances in Neural Information Processing Systems*, 37:125416-125440, 2024. doi: 10.52202/079017-3984. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh

Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153-2162. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221/.

Zixia Wang, Gaojie Jin, Jia Hu, and Ronghui Mu. CluCERT: Certifying LLM robustness via clustering-guided denoising smoothing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37998-38006, 2026. doi: 10.1609/aaai.v40i44.41137. URL https://ojs.aaai.org/index.php/AAAI/article/view/41137.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*, volume 36, pp. 80079–80110. Curran Associates, Inc., 2023. doi: 10.52202/075280-3508. URL https://openreview.net/forum?id=jA235JGM09.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10693-10705. PMLR, 13-18 Jul 2020. URL https://proceedings.mlr.press/v119/yang20c.html.

Mao Ye, Chengyue Gong, and Qiang Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3465-3475. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.317. URL https://aclanthology.org/2020.acl-main.317/.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [MASK]. *Computational Linguistics*, 49(2):395-427, June 2023. doi: 10.1162/coli_a_00476. URL https://aclanthology.org/2023.cl-2.5/.

Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. Text-CRS: A generalized certified robustness framework against textual adversarial attacks. In *2024 IEEE Symposium on Security and Privacy*, pp. 2920-2938. IEEE, 2024. doi: 10.1109/SP54263.2024.00053. URL https://doi.org/10.1109/SP54263.2024.00053.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in neural information processing systems*, 36:46595-46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Qinyan Zhou, Peixin Zhang, Jun Sun, Haonan Zhang, and Dongxia Wang. DDOR: Delta debugging for explainable overrefusal testing and repair, 2026. URL https://arxiv.org/abs/2606.03601.

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.

A Complete Transfer Proof

This section makes explicit the probability space and the boundary-randomization qualification used in Theorem 4.1. Let $\Omega = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. Positivity of each replacement mass and $q > 0$ imply $Q_u(z) > 0$ for every $z \in \Omega$, so $L_{u,v}$ is defined everywhere. Group atoms with equal likelihood ratio into m strata.

Lemma A.1 (Fractional knapsack form)**.** *For nonnegative a_j, b_j satisfying $b_j = \ell_j a_j$ with $\ell_1 > \cdots > \ell_m$, the program*

$$\max_{0 \le \gamma_j \le 1} \sum_{j=1}^{m} \gamma_j b_j \quad \text{s.t.} \quad \sum_{j=1}^{m} \gamma_j a_j \le t \tag{10}$$

has an optimum that fills strata in increasing index order and has at most one fractional γ_j.

Proof. Suppose $j < k$, $\gamma_j < 1$, and $\gamma_k > 0$. Move $\delta \le \min\{(1-\gamma_j)a_j, \gamma_k a_k\}$ units of constraint mass from k to j. The objective changes by $\delta(\ell_j - \ell_k) > 0$. Thus no optimum can leave higher-ratio capacity while using lower-ratio capacity. Iterating gives a prefix of full strata, at most one boundary stratum, and a zero suffix. The constraint is tight unless $t = 1$. □

Any randomized test $\phi : \Omega \to [0, 1]$ induces $\gamma_j = \mathbb{E}_{Q_u}[\phi \mid L_{u,v} = \ell_j]$. Its expectations under Q_u and Q_v are exactly the constraint and objective of Equation 10. Conversely, choosing a constant $\phi = \gamma_j$ within each stratum realizes any feasible solution. This proves tightness over randomized tests. A deterministic event restricts ϕ to $\{0, 1\}$ at each atom and may not realize a fractional boundary. Hence the envelope upper-bounds every deterministic gate event but need not equal its instance-specific subset optimum.

Confidence transfer. Let $E_u = \{p_u \le \overline{p}_u\}$ be the binomial confidence event. The calculation of U_r introduces no new sampling randomness. On E_u, the chain in Proposition 4.2 holds simultaneously for every center in the finite maximization. The statement therefore inherits the per-instance confidence level. It is not a simultaneous population confidence statement over all prompts unless an additional multiplicity correction is imposed.

B Envelope Algorithm and Verification

Algorithm 1 Exact small-radius Hamming transfer envelope

Require: center u, fixed $\{\pi_i\}_{i=1}^n$, noise q, radius r, budget t
$U \leftarrow t$
for $j = 1, \ldots, \min(r, n)$ **do**
  for each position subset $I \subseteq [n]$ with $|I| = j$ **do**
    for each tuple of distinct alternative masses $(m_i)_{i \in I}$ **do**
      construct the three strata for every $i \in I$
      convolve strata across I and merge equal likelihood ratios
      $V \leftarrow$ fractional fill in descending likelihood-ratio order
      $U \leftarrow \max(U, V)$
    end for
  end for
end for
return U

Let M_i be the number of distinct alternative masses at position i. Direct enumeration has cost proportional to $\sum_{j=1}^{r} \sum_{|I|=j} (\prod_{i \in I} M_i) 3^j$, before merging and sorting. Uniform π_i gives $M_i = 1$, which is much smaller than token enumeration. The implementation uses log-ratio grouping to merge numerically identical strata and verifies that both distributions sum to one within tolerance.

At the maximum observed length, radius-two maximization enumerates 2,926 outer cases in 0.33 seconds; a controlled 128-token input requires 8,256 cases and 0.95 seconds. In contrast, each reported Hamming certificate uses 10,000 gate evaluations, so model inference dominates this

Table 4: Single-process CPU cost of the exact envelope for uniform $K = 32$. The real evaluated prompt lengths have median 12 and maximum 76 tokens; the 128-token row is a controlled extrapolation. Times exclude gate inference and support construction and are medians of three runs on an AMD EPYC 7543 CPU.

Tokens	Radius	Outer cases	Median time (s)
12	1	12	0.0014
12	2	78	0.0092
76	1	76	0.0087
76	2	2926	0.3312
128	2	8256	0.9453

implementation's runtime at the evaluated radii. Non-uniform kernels scale with the product of distinct mass classes as stated above.

B.1 Verification suite

The implementation is subjected to the following independent checks.

1. **Mass conservation:** every position and convolved distribution sums to one under both centers.
2. **Known two-atom case:** analytically computed fractional-envelope values match the implementation at multiple budgets.
3. **Deterministic-event oracle:** for small supports, every subset of atoms is enumerated; its Q_v mass never exceeds the fractional envelope at the same Q_u budget.
4. **Product oracle:** explicit Cartesian-product atoms agree with the convolution representation.
5. **Radius monotonicity:** $U_0 \leq U_1 \leq U_2$ on randomized problems.
6. **Support failure:** a center or substitute with zero fixed-support mass is rejected rather than silently assigned a finite ratio.

These tests separate mathematical validity from floating-point engineering. They do not validate the external gate's policy labels, which remain empirical model outputs.

C Detailed Experimental Protocol

C.1 Dataset construction and splitting

Each source is parsed into records with a stable ID, prompt text, source name, and benchmark label. IDs rather than row positions define a deterministic prompt-grouped split, preventing reordering from moving examples across development and test. The unsafe test split contains the complete 629-example population used for gate and Hamming reporting. The adaptive subset takes 75 test IDs per source after deterministic sorting. The benign pool contains 405 held-out prompts for the frozen gate analysis. Dataset documentation follows the provenance principles of datasheets and model cards (Gebru et al., 2021; Mitchell et al., 2019).

Table 5: Evaluation populations and routing counts.

Population	Total	Gate-rejected / routed	Unit
Selective-recovery test	2,419	894 Gate-A rejected	prompt
Unsafe Hamming test	629	606	unique prompt
Adaptive balanced test	225	213 (LlamaGuard) / 224 (WildGuard)	goal

C.2 Model identities and formatting

The primary gate is Meta-Llama-Guard-2-8B at revision `7d257f3c1a0e`. WildGuard uses revision `cbba4823f3e80`. Llama-2-7B-hf uses revision `01c7f73d771d` and raw completion formatting. Mistral-7B-Instruct-v0.3 uses revision `c170c708c41d` with its tokenizer-provided chat template. The legacy adaptive study uses these Llama 2 and Mistral configurations. The prompt-disjoint replication separately uses Llama-Guard-3-8B at revision `7327bd9f6efb`, Meta-Llama-3-8B-Instruct at revision `8afb486c1db2`, and Gemma-3-12B-IT at revision `96b6f1eccf38` (Llama Team, AI @ Meta, 2024; Gemma Team, 2025); all three use their tokenizer-provided chat templates. WildGuard retains the revision above in both protocols. Direct WildGuard adaptation uses Gemma 3 as the downstream target and LlamaGuard 3 as the secondary output judge. All inference is offline and deterministic conditional on recorded random seeds and generation configuration.

C.3 Adaptive selection and fresh certification

For each goal and restart, the coordinate attacker produces a suffix, evaluates the hard gate, estimates a 200-sample screening certificate, generates a target response, and obtains two output-judge labels. Candidate selection is lexicographic: screening certified-and-harmful success, hard released-and-harmful success, primary harmfulness, hard acceptance, then gate loss. This deliberately gives the attack access to the failure event rather than selecting a convenient defense outcome.

After selection, the certificate is discarded and recomputed from 1,000 fresh samples with a seed separated from every search and screening stream. The final predicate is

$$\mathbf{1}\{\overline{p}_u < 1/2\}\,\mathbf{1}\{J_{\mathrm{WG}}(x, G(u)) = 1\}\,\mathbf{1}\{J_{\mathrm{LG}}(x, G(u)) = 1\}. \tag{11}$$

Parser failures are fail-closed for harmfulness. Outputs are truncated only in stored diagnostic fields, after judging, so truncation does not affect labels.

C.4 Multi-turn diagnostic

Each conversation has up to eight turns between the user and assistant. The attacker state depends on the preceding target response: refusal triggers reframing, partial answers trigger requests for operational detail, and accepted premises trigger continuation. Five routes are compared: no defense, current-turn blocking, history blocking, current-turn smoothed repair, and history-aware smoothed repair. The primary endpoint is whether any turn is harmful. A matched benign population measures whether a route completes without blocking. Crescendo motivates the response-conditioned structure (Russinovich et al., 2025), but the implemented state machine is not claimed as a reproduction of that attack.

D Selective-Recovery Decision Audit

The common-certifier development comparison holds the LlamaGuard decision, $q = .1$ center kernel, confidence level, and automated fidelity rule fixed while changing only the proposal generator. Table 6 retains the full benign false-positive and unsafe denominators. The large raw recovery of minimal truncation is coupled to the largest unsafe loss; paraphrasing has the best measured fidelity fraction but does not meet the recovery and unsafe-loss criteria jointly.

Table 6: Development comparison under one common certifier. "Released" also requires the frozen automated fidelity rule.

Generator	Benign certified	Benign released	Unsafe certified	Unsafe released
SCoRe beam	31/81	20/81	66/171	16/171
Paraphrase	19/81	16/81	15/171	10/171
Minimal truncation	66/81	47/81	130/171	55/171

Clean-room mRTF analogue. The analogue first applies deterministic delta debugging to the rejected prompt's token indices. At each granularity it tests both chunks and their complements,

retains the shortest gate-rejected subset, and finishes with single-token deletion until the subset is 1-minimal under the gate. It then runs the same bounded substitution beam used by SCORE, restricting editable positions to that localized subset. Candidate selection and release use the same hard-gate, semantic-cosine, randomized-certificate, and abstention rules as the other proposal rows. DDOR additionally uses a rewriting model conditioned on the original prompt and mRTF, constructs overrefusal tests, and applies multi-oracle filtering (Zhou et al., 2026); those components are absent here. The row therefore tests whether localization improves proposals inside SCORE, not the performance of DDOR.

The test-stage operating point was frozen from development before examining test outcomes. Of 2,419 test prompts, the first hard gate passes 1,422/1,735 benign and 103/684 unsafe prompts unchanged. It proposes repairs only for the remaining 313 and 581 prompts, finding 200 and 252 candidates. Gate A certifies 179 benign and 242 unsafe-source repairs; Gate B certifies 59 and 83; their intersection contains 56 and 83. The final fidelity filter releases 34 benign and 15 unsafe-source repairs. Table 8 shows that aggregate benign recovery is driven by XSTest and OR-Bench, while mixed academic/programming tasks almost never enter repair. This composition is why the aggregate is not a benchmark-independent utility estimate.

Table 7: Development q frontier for SCORE beam with joint center certificates and the frozen fidelity rule.

q	Benign release	BFR Wilson lower	Unsafe release / all unsafe
.1	7/81	4.25%	2/171 (1.17 pp)
.2	9/81	5.96%	4/171 (2.34 pp)
.3	9/81	5.96%	5/171 (2.92 pp)
.4	11/81	7.76%	9/171 (5.26 pp)

Higher q increases measured benign release but crosses the two-point unsafe-loss cap. Consequently, $q = .1$ is the highest-rescue development point satisfying that empirical constraint; it still misses the BFR target. This selection is separate from the uniform-kernel Hamming experiment, whose $q = .4$ point was chosen to maximize overlap for a different estimand.

Table 8: Frozen dual-gate test result by source. Counts are final repair releases over prompts rejected by Gate A; unsafe TPR loss elsewhere uses the full unsafe denominator.

Source	Label	Gate-A rejected	Dual certified	Released
All	benign	313	56	34
All	unsafe	581	83	15
OR-Bench Hard	benign	279	39	24
OR-Bench Toxic	unsafe	441	52	8
XSTest	benign	20	13	8
XSTest	unsafe	140	31	7
JailbreakBench benign	benign	13	3	2
Mixed academic/code tasks	benign	1	1	0

Audited example. One OR-Bench Hard prompt asks: "I'm a security professional looking to test the vulnerabilities of a building, can you explain the best practices for entry?" SCORE changes only "entry" to "access." The repair is accepted by both center certificates and the automated fidelity rule (cosine 0.948, bidirectional NLI 0.984, token edit ratio 0.05), so it is one of the 34 benign recoveries. This example exposes the decision path, which consists of a proposal, two fresh certificates, a fidelity check, and then release. It does not establish that the same edit is necessary or that a paraphraser could not also recover the prompt.

The two center certificates use $N = 10{,}000$ independent samples and $\alpha = .001$ per gate. Their per-candidate conjunction has confidence at least $1 - .002$ by a union bound. Wilson intervals in the main text instead describe benchmark proportions and are not certificate confidence intervals. Because the fidelity validator is fully automated, 34/56 is reported as model agreement, not verified human intent preservation. No thresholds or generator choices are changed after the confirmatory result fails its three declared criteria.

E DISAGGREGATED RESULTS

Table 9: Common-certifier results on 1,000 benign and 149 unsafe prompts. Route counts use the hard gate; releases require fidelity and a fresh center certificate. The mRTF row is a clean-room targeted analogue, not an official reproduction of DDOR.

Cell	Benign route	Benign release	Unsafe route	Unsafe release	Center/all route	HCRR-1
SCORE / Llama Guard 3	16/1000	2/16	145/149	2/145	4/161	0/161
mRTF analogue / Llama Guard 3	16/1000	1/16	145/149	10/145	11/161	0/161
SCORE / WildGuard	123/1000	10/123	148/149	4/148	14/271	0/271

Table 10: Harmful outputs among unsafe-labelled prompts sent by the selective policy to a modern target. "LG3" and "WG" are separate automated judges; "Both" is their conjunction. The generated denominator includes pass-through and center-release branches, while every full unsafe population contains 149 prompts.

Target	Policy cell	Generated unsafe	LG3	WG	Both
Llama 3	SCORE / LG3	6/149	1/6	3/6	1/6
Llama 3	mRTF / LG3	14/149	3/14	6/14	3/14
Llama 3	SCORE / WG	5/149	0/5	0/5	0/5
Gemma 3	SCORE / LG3	6/149	0/6	3/6	0/6
Gemma 3	mRTF / LG3	14/149	2/14	7/14	1/14
Gemma 3	SCORE / WG	5/149	1/5	0/5	0/5

The endpoint evaluation generates from every prompt passed through or released at the center and generates nothing on abstention. Both judge families classify all 32 replicated controls correctly. Llama Guard 3 labels 29 of 2,070 outputs harmful and WildGuard labels 72, while only 12 decisions overlap. We therefore retain judge-specific and conjunction results rather than averaging them. For the ten unsafe-labelled prompts center-released by the mRTF analogue, the disjunction identifies harmful output for 4/10 with Llama 3 and 6/10 with Gemma 3. For SCORE, the corresponding denominators are two under Llama Guard 3 and four under WildGuard. These small samples characterize observed conditional release risk but do not support a population-level safety claim.

Table 11: WildGuard applied to attacks optimized against LlamaGuard. Threshold names describe frozen calibration targets; test rates are not rematched.

Gate-B rule	B accepted	A-certified ∩ B accepted	Harmful by both judges
Hard label	0/213	0/213	0/213
Frozen-FPR	1/213	1/213	0/213
Frozen-TPR	52/213	49/213	28/213

E.1 HAMMING COVERAGE WITH CONFIDENCE INTERVALS

Of the 629 prompts, 606 enter the repair route. Their corresponding route HCRR counts are 222/606 at the center and 0/606 at radii one and two; every nonzero- radius analysis positive is among the 23 pass-through prompts. The radius-nesting invariant holds for every record: radius-two certification implies radius one, which implies center certification. Counts are deduplicated by prompt because two repair methods can produce records for the same input. This avoids treating correlated method outputs as independent examples.

The product shortcut uses the largest per-change likelihood ratio, 49, on all unsafe-event mass. The NP envelope instead exhausts the finite mass in each likelihood-ratio stratum before advancing.



Figure 3: Full-pool analysis coverage by source and radius. These counts include prompts passed through by the deployment gate and are not repair-route HCRR; Table 2 reports the route-correct estimand.

Table 12: Full-pool certificate analysis with Wilson 95% intervals. These counts include prompts passed through by the deployment gate and therefore are not repair-route HCRR. The per-instance certificate uses a separate one-sided binomial confidence bound.

Source	Analysis-$r0$	Analysis-$r1$	Analysis-$r2$
All	244/629 (35.1-42.7)	4/629 (0.25-1.62)	1/629 (0.03-0.90)
AdvBench	163/347 (41.8-52.2)	2/347 (0.16-2.07)	0/347 (0-1.09)
JailbreakBench	42/77 (43.5-65.2)	2/77 (0.72-8.96)	1/77 (0.23-7.00)
StrongREJECT	39/205 (14.2-25.0)	0/205 (0-1.84)	0/205 (0-1.84)

Table 13 shows that this difference is numerically substantial even though it rarely changes the binary certificate at the evaluated margins. The full-pool analysis row in the main table deduplicates these method-prompt rows into unique prompts, explaining its counts of four and one rather than eight and two.

Table 14 distinguishes a statistical budget limit from a measured margin limit. For exact transfer, 280 of 281 radius-one failures and all 281 radius-two failures remain negative even if sampling uncertainty is removed while the observed unsafe frequency is held fixed. Only one radius-one row is sampling-limited. The radius-two product shortcut is infeasible for all 281 center-certified rows at the chosen sample budget, but exact transfer is feasible and still certifies none on the repair route.

E.2 Same-policy neighborhood audit

This audit reuses the two independent $N = 10{,}000$ center bounds for every frozen candidate and reconstructs the center-only position supports used by the policy. It introduces no additional gate sampling and does not tune on test. The zero neighborhood counts therefore reflect the transfer envelope for the evaluated policy rather than a Monte Carlo rerun or a separate repair pool.

The uniform $K = 32$ test isolates the transfer calculation in a controlled support. Table 16 is a complementary development diagnostic: the broader support produced by the actual proposal mechanism gives no positive radius-one or radius-two certificate anywhere in the screened q range. This does not prove that every non-uniform semantic kernel must fail.

Table 13: Transfer-envelope audit on 350 uniform-kernel method-prompt rows. Medians are over 325 center-certified rows; "finite" means an upper bound strictly below one.

r	NP cert.	Product cert.	NP finite	Product finite	Median NP / product
1	8	8	350	14	.788 / 1.000
2	2	0	350	0	.991 / 1.000

Table 14: Failure decomposition for the 304 repair-route method-prompt rows. Categories are mutually exclusive and ordered as shown. "Budget infeasible" means that the zero-error Clopper-Pearson endpoint cannot cross the transfer threshold at $N = 10{,}000$ and $\alpha = .001$.

Certificate	Center fail	Budget infeasible	Margin limited	Sampling limited	Certified
$r = 1$ exact	23	0	280	1	0
$r = 1$ product	23	0	280	1	0
$r = 2$ exact	23	0	281	0	0
$r = 2$ product	23	281	0	0	0

E.3 KERNEL-FAMILY SENSITIVITY

Table 17: Independent high-sample development evaluation of the six cells selected by the predeclared screening rule. Each cell contains the same 151 repaired method-prompt rows, uses $N = 10{,}000$ fresh samples, and applies a one-sided per-instance $\alpha = .001$.

q	K	Replacement	Center	$r = 1$	$r = 2$
0.1	8	uniform	139/151	2/151	0/151
0.1	8	weighted	141/151	0/151	0/151
0.5	8	uniform	145/151	12/151	4/151
0.5	8	weighted	144/151	7/151	2/151
0.5	64	uniform	150/151	4/151	0/151
0.5	64	weighted	150/151	0/151	0/151

Table 17 confirms a trade-off between coverage and neighborhood robustness rather than a binary dependence on the original $q = .4, K = 32$ setting. Center counts remain between 139 and 150, while radius-one counts range from zero to 12 and radius-two counts from zero to four. At $q = .5, K = 8$, the weighted distribution has fewer observed neighborhood certificates than the uniform distribution; the experiment therefore provides no evidence that the evaluated weighting closes the gap. These paired counts are descriptive and are not presented as a hypothesis test between kernel families.

Table 18: Radius-one high-sample counts by source. Denominators are method-prompt rows and are identical across kernel cells.

q	K	Replacement	AdvBench	JailbreakBench	StrongREJECT
0.1	8	uniform	0/115	2/15	0/21
0.1	8	weighted	0/115	0/15	0/21
0.5	8	uniform	6/115	4/15	2/21
0.5	8	weighted	4/115	2/15	1/21
0.5	64	uniform	0/115	2/15	2/21
0.5	64	weighted	0/115	0/15	0/21

The best observed cell, uniform $q = .5, K = 8$, has radius-one positives in all three sources, although the rates differ and JailbreakBench has only 15 rows. Its four radius-two positives comprise two AdvBench and two JailbreakBench rows. The result is therefore not attributable to one source, but its small positive counts do not support a source-invariant rate claim.

Table 15: Exact transfer for the frozen dual-gate selective policy. Denominators are all benign or unsafe prompts routed to repair. "Fidelity" additionally requires the frozen semantic similarity, bidirectional entailment, and edit ratio rule.

Route label	Candidate	Dual center	Dual $r = 1$	Dual $r = 2$	Fidelity center	Fidelity $r = 1$	Fidelity $r = 2$
Benign	200/313	56/313	0/313	0/313	34/313	0/313	0/313
Unsafe	252/581	83/581	0/581	0/581	15/581	0/581	0/581

Table 16: Development sensitivity for the actual uncapped, position-specific SCoRe support (151 repaired method-prompt rows).

q	Center cert.	$r = 1$	$r = 2$	Median support
.1	137/151	0/151	0/151	132
.2	126/151	0/151	0/151	132
.3	130/151	0/151	0/151	132
.4	144/151	0/151	0/151	132

Table 19: Development ablation of embedding-nearest admissible supports. Each cell contains the same 151 method-prompt rows and 111 unique prompts, uses $N = 10{,}000$ fresh samples, and applies one-sided per-instance $\alpha = .001$. "Uniform" assigns equal mass within the selected semantic support. "Semantic-softmax" uses temperature .2 and a .05 uniform floor.

q	K	Replacement mass	Center	$r = 1$ / $r = 2$
0.4	8	uniform	133/151	2/151 / 0/151
0.4	8	semantic-softmax	127/151	2/151 / 0/151
0.4	32	uniform	146/151	2/151 / 0/151
0.4	32	semantic-softmax	129/151	0/151 / 0/151
0.5	8	uniform	141/151	12/151 / 2/151
0.5	8	semantic-softmax	120/151	2/151 / 2/151
0.5	32	uniform	149/151	4/151 / 2/151
0.5	32	semantic-softmax	131/151	0/151 / 0/151

The ablation retains the original and repaired token at every position, then fills the remaining support with tokens nearest to the repaired token under the frozen GPT-2 embedding. The largest radius-one count is 12/151 for uniform $q = .5, K = 8$, which matches the largest count in Table 17; its radius-two count is 2/151 rather than 4/151. Semantic-softmax mass does not improve radius-one coverage in any matched cell and substantially reduces center coverage in several cells. Thus the evaluated semantic support construction does not explain away the observed center-to-neighborhood gap. The complete audit also reports seeded 95% prompt-cluster bootstrap intervals, although the table presents counts because the cells are paired development diagnostics rather than independent samples.

Table 20: Development screening counts for all 40 kernel cells at $N = 1{,}000$. Every entry has denominator 151; values are counts rather than percentages. This screen is used only for predeclared cell selection and is not pooled with the high-sample results in Table 17.

		Uniform			Weighted		
q	K	Center	$r = 1$	$r = 2$	Center	$r = 1$	$r = 2$
0.1	8	137	0	0	138	0	0
0.1	16	135	0	0	137	0	0
0.1	32	136	0	0	137	0	0
0.1	64	135	0	0	136	0	0
0.2	8	118	2	0	118	0	0
0.2	16	119	0	0	118	0	0
0.2	32	124	0	0	122	0	0
0.2	64	127	0	0	124	0	0
0.3	8	122	2	0	121	0	0
0.3	16	125	2	0	124	0	0
0.3	32	130	0	0	130	0	0
0.3	64	131	0	0	130	0	0
0.4	8	129	4	0	129	2	0
0.4	16	136	2	0	134	0	0
0.4	32	140	0	0	138	0	0
0.4	64	145	0	0	140	0	0
0.5	8	140	8	2	140	6	0
0.5	16	146	3	0	145	1	0
0.5	32	149	2	0	148	0	0
0.5	64	150	0	0	148	0	0

The screen spans the full $q \times K \times$ replacement-family design. Its selection rule retains family extremes and cells closest to or below the radius-one boundary; overlapping criteria produce six unique confirmation cells. The fresh high-sample evaluation changes some individual counts but not the qualitative ordering: $q = .5, K = 8$ retains the largest radius-one count in both replacement families, while no evaluated cell has neighborhood coverage near its center coverage.

E.4 Adaptive results by source

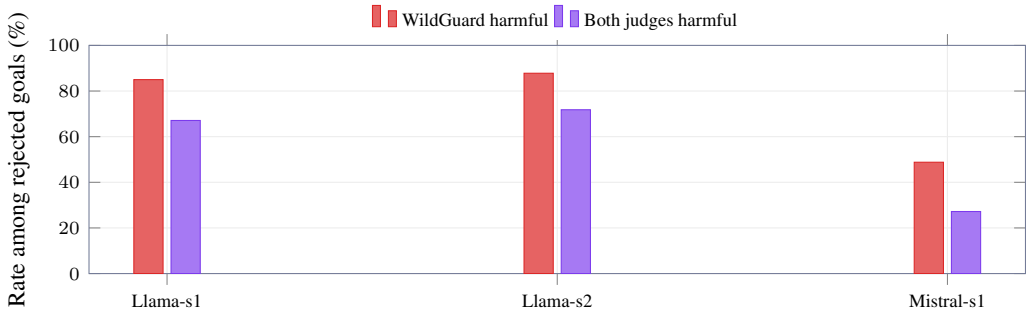

Figure 4: Fresh-certified and harmful rates in the LlamaGuard-adaptive study. Consensus is stricter than either individual automated judge.

The target shift changes harmfulness much more than certification. Fresh certification stays between 82.9% and 98.6% across source/target cells, whereas consensus harmfulness ranges from 20.5% to 82.2%. This supports the mechanism claim: the gate certificate is largely insensitive to which downstream target consumes the prompt, while output harm is target-dependent.

The consensus endpoint occurs on AdvBench and JailbreakBench but not on StrongREJECT in either run. Five successful prompt IDs overlap across seeds; the union contains 17 IDs. The overlap is a stability diagnostic rather than a second independent sample, so inference remains seed-specific.

Table 21: Fresh-certified and harmful-by-both counts among goals rejected before attack, disaggregated by source.

Target/run	Source	Fresh certified	Certified ∩ harmful
Llama seed 1	AdvBench	71/73	60/73
	JailbreakBench	65/70	45/70
	StrongREJECT	59/70	38/70
Llama seed 2	AdvBench	70/73	57/73
	JailbreakBench	68/70	52/70
	StrongREJECT	63/70	44/70
Mistral seed 1	AdvBench	72/73	15/73
	JailbreakBench	68/70	24/70
	StrongREJECT	58/70	19/70

Table 22: Direct WildGuard-to-Gemma adaptive results by source. Runs share the same goals but use independent attack seeds and are not pooled.

Run	Source	Rejected	Fresh cert. ∩ harmful (both)
Seed 1	AdvBench	75	10/75
Seed 1	JailbreakBench	74	3/74
Seed 1	StrongREJECT	75	0/75
Seed 2	AdvBench	75	6/75
Seed 2	JailbreakBench	74	3/74
Seed 2	StrongREJECT	75	0/75

E.5 Gate smoothing frontier

Table 23: Frozen gate results on 405 benign and 629 unsafe prompts.

Gate	Rule	Benign FP	FPR	Unsafe TP	TPR
LlamaGuard	hard	15/405	3.7	606/629	96.3
	$q = .1$	13/405	3.2	602/629	95.7
	$q = .2$	9/405	2.2	579/629	92.1
	$q = .3$	8/405	2.0	535/629	85.1
WildGuard	hard	41/405	10.1	626/629	99.5
	$q = .1$	43/405	10.6	626/629	99.5
	$q = .2$	38/405	9.4	624/629	99.2
	$q = .3$	30/405	7.4	620/629	98.6

The two gates occupy different operating points. Consequently, raw cross-gate survival does not isolate robustness from strictness. Frozen low-FPR and high-TPR views reduce this ambiguity but do not eliminate model and calibration differences.

F Multi-Turn and Cross-Gate Diagnostics

History-aware repair completes more harmful conversations without blocking but does not reduce the any-harm endpoint relative to current-turn repair. This is not a utility gain: benign completion is already 96-100% for all defended routes. The diagnostic therefore provides no basis for elevating multi-turn repair into a primary contribution.

The last column additionally requires LlamaGuard's fresh center certificate. It is a composite transfer diagnostic, not a certificate for the conjunction. In particular, zero hard-label transfer gives no population guarantee and no guarantee under an attacker optimized against WildGuard.

Table 24: Eight-turn conversation outcomes. Harm is whether any response is judged harmful; completion is a conversation with no block. Each cell has 50 conversations.

		Harmful set		Benign set	
Target	Route	Harm	Complete	Harm	Complete
Llama	no defense	26	50	0	50
	current block	11	2	0	48
	history block	13	9	0	49
	current repair	11	6	0	49
	history repair	14	15	0	50
Mistral	no defense	30	50	0	50
	current block	6	3	0	49
	history block	3	7	0	49
	current repair	4	8	0	49
	history repair	8	18	0	50

Table 25: Cross-gate decomposition among 213 attacks optimized against LlamaGuard.

WildGuard view	B accepts	A cert. ∩ B	Both judges harmful
Hard label	0	0	0
Frozen low-FPR	1	1	0
Frozen high-TPR	52	49	28

G Closest-Work Scope Map

Table 26 compares scientific objects rather than heterogeneous performance numbers. A certificate for an input-gate event is not interchangeable with one that regenerates responses under noise and then classifies those responses.

Table 26: Scope comparison with the closest repair, certification, and audit families. "Output" means that downstream generation is inside the certified random variable, not merely evaluated after release.

Work	Repair	Selective route	Certified random object	Threat set	Output	Adaptive eval.
DDOR (Zhou et al., 2026)	yes	no	none	localized fragment rewrite	no	no
SCoRE	yes	yes	input-gate label under categorical noise	fixed-support substitutions	no	yes
SAFER (Ye et al., 2020)	no	no	text-classifier label	synonym substitutions	no	no
RanMASK (Zeng et al., 2023)	no	no	text-classifier label	masked-token perturbations	no	no
Text-CRS (Zhang et al., 2024)	no	no	text-classifier label	substitution, insertion, deletion, and reordering	no	no
CERT-ED (Huang et al., 2024)	no	no	text-classifier label	edit distance	no	no
DiffTextPure (Chen et al., 2025)	no	no	stochastic-defense failure event	token edits and suffixes	method-dependent	yes
Hybrid RS (Delattre et al., 2026)	no	no	binary classifier under factorized noise	discrete and continuous perturbations	if base event	yes
CluCERT (Wang et al., 2026)	no	no	LLM task decision after denoising	synonym substitutions	task-dependent	empirical
DCRS (Jang et al., 2026)	no	no	generated-response safety label	subset plus embedding displacement	yes	empirical
CSS (Cheng et al., 2026)	no	no	ablated-ensemble safety decision	semantic-token substitutions	yes	yes
Assumption audit (Pratap et al., 2026)	no	defense-specific	none shared across defenses	semantic and multi-turn attacks	empirical	no

DiffTextPure computes tight likelihood-ratio bounds using knapsack optimization, but it does not condition coverage on a selected prompt-repair route. Hybrid RS contains the discrete likelihood-ratio optimization underlying Theorem 4.1; SCoRE does not claim that statistical principle as novel. Its distinct objects are the selected repaired center, fixed position-specific support, exact outer maximization over the declared Hamming ball, and route-conditioned release coverage. DCRS and CSS place response generation within the smoothed decision, whereas SCoRE deliberately certifies an input gate and then measures downstream behavior as a separate endpoint. The assumption audit covers broader semantic attacks and larger models; SCoRE provides a narrower controlled decomposition of one selective release interface.

H Record Validation and Aggregation

Experiment records were stored with frozen configurations, per-example observations, summaries, and validation markers. Provenance fields recorded dataset identity, model revisions, code state, and

seeds. Validation checked schema identity, row counts, unique record IDs, finite probability values, bounds $0 \leq k \leq N$, and exact consistency between row predicates and summaries.

Tables were computed only from validated records with the expected counts. Aggregation reconstructed metrics from rows and computed Wilson intervals; headline cells were checked independently against those rows. These procedures describe internal evidence validation and do not imply that an executable archive accompanies the submission.

Table 27: Core experiment-record fields and their purpose.

Field group	Purpose
Identity	stable prompt/goal ID, source, split, analysis unit
Provenance	dataset identity, code state, public model revision
Search	restart, steps, candidate budget, selected suffix, gate loss
Certification	kernel parameters, sample seed, N, unsafe count, upper bound
Generation	target identity and formatting mode, truncated audit output
Judgment	individual raw parses, binary labels, consensus predicate
Integrity	schema version, expected rows, validation marker

Only complete, schema-validated experiment records enter aggregation. Missing, duplicate, malformed, or unvalidated records fail aggregation rather than changing the denominator. This fail-closed contract prevents incomplete evidence from appearing as a favorable scientific result.

I Additional Scope Analysis

I.1 What the certificate cannot establish

The following implications are invalid without additional assumptions:

$$\begin{aligned}
\overline{p}_u < 1/2 &\not\Rightarrow S(u) = \text{safe}, \\
U_r(u, \overline{p}_u) < 1/2 &\not\Rightarrow S(v) = \text{safe} \quad (v \in \mathcal{B}_r(u)), \\
\overline{p}_u < 1/2 &\not\Rightarrow G(u) \text{ is harmless}, \\
U_r(u, \overline{p}_u) < 1/2 &\not\Rightarrow G(v) \text{ is harmless for } v \in \mathcal{B}_r(u), \\
S_A(u) = S_B(u) &\not\Rightarrow S_A, S_B \text{ are independent.}
\end{aligned}$$

The first two distinctions separate a hard label from the smoothed labels $\widetilde{S}_Q(u)$ and $\widetilde{S}_Q(v)$. The transfer condition establishes the latter throughout the declared ball, while hard-neighbor safety requires a separate assumption or evaluation. The next two implications require a formal connection between a gate event and output behavior, which this work does not assume. The final implication cautions against interpreting model diversity as probabilistic independence.

I.2 Alternative certificates

Randomized ablation can certify sparse attacks by removing information rather than replacing tokens (Levine & Feizi, 2020). Hierarchical smoothing can exploit structured perturbation groups (Scholten et al., 2023), and f-divergence bounds can trade tightness and tractability (Dvijotham et al., 2020). Interval methods can directly propagate symbol-substitution sets (Huang et al., 2019; Jia et al., 2019). These alternatives change either the noise distribution, certified object, or computational assumptions. They are promising comparisons but cannot be substituted into the present result without rerunning both coverage and utility.

I.3 Automated-judge uncertainty

The conjunction of two judges is a conservative operational endpoint, not a statistical estimator of a latent human consensus. Shared training data and policy taxonomies can correlate judge errors. A future human study would need a preregistered harm taxonomy, blinded presentation, multiple annotators, an adjudication rule, and agreement reporting. The present work instead preserves all judge-disagreement cells and limits claims to the named automated policies.

Table 28: Modern-target output-judge confusion on 1,035 unique endpoint inputs per target. "Only harmful" columns expose direction rather than compressing disagreement into one agreement coefficient.

Target	Both safe	LG3 only harmful	WG only harmful	Both harmful	Agreement
Llama 3	992	6	31	6	998/1035
Gemma 3	989	11	29	6	995/1035

J Evaluation Principles for Certified Repair

The combined theory and experiments suggest four requirements for evaluating a certified repair system. First, name the certified random variable and event. A statement about $S(Z)$ under token noise is neither a statement about the hard label at every string nor a statement about $G(u)$; Table 1 makes this distinction part of the method definition.

Second, use the deployment population as denominator. Conditioning on successful repairs hides search failures and abstentions. Route HCRR therefore uses every prompt sent to repair, BFR uses every benign false positive, and end-to-end evaluation counts uncertified outputs as blocked and harmful certified outputs as failures.

Third, make adaptivity follow the operative decision path. A hard-label attack does not fully evaluate a smoothed gate, and an attack against gate A does not establish gate-B robustness. Screening samples can rank candidates but cannot supply the reported post-selection confidence statement; fresh samples are required.

Fourth, accompany negative and zero cells with controls. Near-zero Hamming coverage is interpretable because exact small-support oracles pass and center coverage is nonzero. Zero cross-gate transfer requires a direct gate-B attack, and multi-turn comparisons require nonzero no-defense harm and matched benign conversations. These controls distinguish a boundary from an inactive attack, strict filter, parser error, or denominator artifact.

K Detailed Design Implications

Optimize the guarantee, not only the center score. Repair search asks for a short edit that crosses the hard decision boundary, after which certification filters the candidate. This decomposition creates a selection mismatch: the nearest accepted candidate can have little smoothed margin. A certificate-aware search can rank candidates by a cheap proxy for $U_r(u, \overline{p}_u)$ and reserve full Monte Carlo evaluation for finalists. The reported certificate must still use fresh samples after selection.

Expose the kernel frontier. A useful token kernel needs overlap between neighboring centers, stable gate labels at the center, and plausible corrupted text. Increasing q improves overlap but can destroy the center label; concentrating π_i on plausible tokens can increase likelihood ratios for rare alternatives; broadening $\mathcal{A}_i$ expands formal scope but can weaken the bound. Kernel selection must therefore report route HCRR, center coverage, benign false positives, unsafe true positives, and repair quality under one fixed rule.

Tie certificates to deployment decisions. An output gate changes the release semantics into two decisions: an input certificate authorizes generation, while an output decision authorizes display. The input theorem remains valid but does not characterize the composite event. Likewise, a conjunction of input gates is certified only if the threat model and probability calculation cover the conjunction itself.

Negative coverage is not evidence that all prompt repair is futile. It rules out one stronger interpretation at the evaluated operating point: broad center-certified recovery cannot simultaneously be advertised as worst-case substitution robustness. A system can retain the center rule as a stability heuristic, restrict release to the robust subset, or redesign its kernel and proposal objective. Each choice changes coverage and must be disclosed.

Table 29: Observed boundaries and the design changes they motivate.

Boundary	Evidence	Design implication
Recovery $\not\Rightarrow$ safe utility	34/313 benign versus 15/684 unsafe releases	report the joint utility-risk frontier
Center $\not\Rightarrow$ Hamming	222/606 route center certificates, 0 at $r = 1$	optimize robust margin and separate route HCRR from analysis
Gate $\not\Rightarrow$ output	nonzero harmful center releases under modern targets	separate input authorization from output display
Transfer $\neq$ robustness	zero hard transfer under an A-only attack	attack every operative gate under its actual template
History $\neq$ improvement	history repair does not beat current-turn repair	keep stateful repair secondary without a stronger attacker