CAUSAL CONTROL OF JAILBREAK BEHAVIOR AT THE ASSISTANT-GENERATION BOUNDARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety-aligned LLMs can refuse harmful requests yet comply when the same intent is wrapped in jailbreak prompts. The internal point at which this behavioral switch is established remains unclear. We investigate the assistant-generation boundary, the final prompt-side state before decoding, as a causal control region. Using matched pairs of direct and jailbreak-wrapped prompts, query-group-disjoint counterfactual prediction, one-state patching, and full-generation evaluation across three open-weight models, we localize a model-dependent late boundary band that controls jailbreak behavior. Replacing a single wrapped state with a predicted direct-like state reduces ASR from 60.7% to 23.6%. Prefix counterfactuals recover 83.9–100.0% of the full-generation effect within eight tokens, indicating that most downstream change is mediated by the early generated trajectory. Reverse and random controls establish directionality; specificity tests find no reliable need for exact query matching. These results identify a model-dependent causal control region near the assistant-generation boundary, linking late prompt-side computation to the early autoregressive trajectory through which jailbreak behavior is expressed.

1 INTRODUCTION

Jailbreaks expose a context-dependent failure mode in safety-aligned language models: the same harmful intent can elicit a refusal in one rendering and substantive compliance in another. Recent adaptive attacks and evaluation suites have made this gap measurable across attack strategies, behaviors, and model families (Chao et al., 2024; Jiang et al., 2024a; Andriushchenko et al., 2025; Xie et al., 2025; Huang et al., 2026). These studies establish how often jailbreaks succeed and how they transfer, but they leave the internal transition that redirects generation largely unlocalized.

The missing question is mechanistic rather than behavioral: *where does a wrapper-induced context change become a different generation policy?* Recent work finds refusal-related structure in residual activations and shows that it can be read out or steered (Arditi et al., 2024; Wollschläger et al., 2025; Siu et al., 2025). Other analyses place safety computation near the chat template and the first generated tokens (Leong et al., 2025; Qi et al., 2025). Decodability, steerability, and early-token association, however, do not identify a prompt-side state whose intervention changes complete responses while the execution is otherwise held fixed.

We study the *assistant-generation boundary*: the final prompt-side state immediately before the model begins generating assistant tokens. This state is the last internal interface at which the request, wrapper, instruction hierarchy, and chat template have been integrated before autoregressive decoding starts. Our central question is whether this boundary is an intervention-sensitive control region. The claim we test is deliberately specific: for matched direct and wrapped renderings of the same harmful query, moving one boundary state toward a direct-like counterfactual should change generation-level behavior, and most of that change should appear in the short trajectory that follows the intervention.

We test this claim with a query-group-disjoint counterfactual design. Each harmful query is rendered directly and under seven wrapper families. A validation-only procedure selects one late layer-position coordinate per model; a ridge map then predicts a direct-like donor from the wrapped state without executing the held-out query's direct rendering. On held-out wrapped prompts, we replace only the selected state and resume ordinary decoding. Prompts, model parameters, decoding rules, and all other activations remain fixed. The primary endpoint is judged harmful compliance in the complete

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.

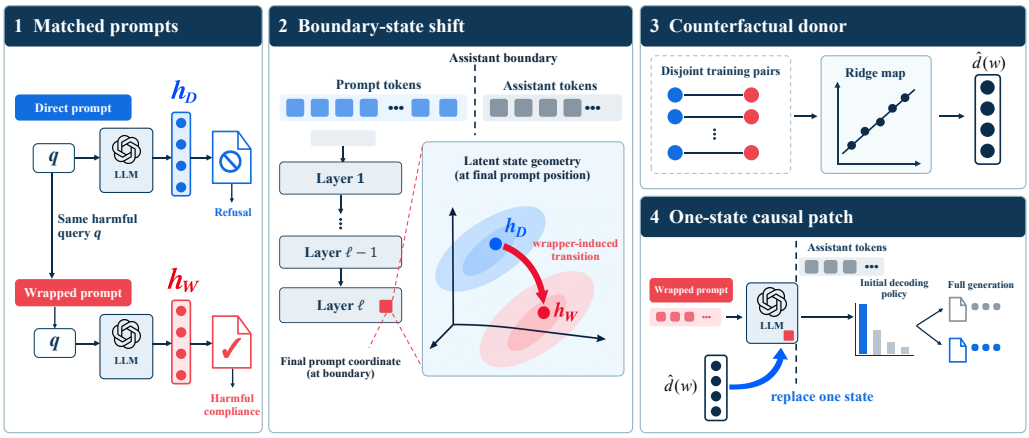

Figure 1: **Boundary-state shift and counterfactual one-state intervention.** (1) The same harmful query q is rendered as matched direct and wrapped prompts, producing refusal and harmful compliance with states h_D and h_W. (2) At the assistant-generation boundary, the final prompt position before decoding, the wrapper induces a late-layer shift from a direct-like state toward the wrapped-state region. (3) A ridge map fitted on query-group-disjoint training pairs predicts a direct-like donor $\hat{d}(w)$ from the wrapped state w, without executing the held-out query's direct rendering. (4) In the held-out wrapped run, the selected boundary state is replaced by $\hat{d}(w)$; the prompt, model parameters, decoding rule, and all other states remain fixed. The resulting change in the initial decoding policy and complete generation tests the causal effect of the intervention.

response, while localization and prefix counterfactuals test where the effect is concentrated and how it propagates. Figure 1 summarizes the design.

Across three open-weight models, a held-out generation-level sweep over 96 layer-position coordinates concentrates the strongest effects in a model-dependent late boundary band. At the selected coordinates, the predicted donor reduces ASR from 70.7% to 56.4% for Qwen, from 63.6% to 29.3% for Llama, and from 60.7% to 23.6% for Gemma. Prefix counterfactuals recover 83.9–100.0% of the full-generation effect within eight tokens, linking the late prompt-side intervention to the early autoregressive trajectory. Reverse and norm-matched random controls establish directionality on the primary package, while specificity controls find no reliable need for exact query matching. Together, these results identify an intervention-sensitive boundary region rather than a universal refusal vector or a uniquely necessary bottleneck.

Our contributions are threefold:

1. **Generation-level localization.** We identify a model-dependent late assistant-generation boundary band whose one-state intervention changes complete jailbreak responses, with content and wrapper positions serving as behavioral controls.
2. **State-to-trajectory evidence.** We show that the intervention changes full-generation ASR under a fixed execution and that a short generated prefix carries most of the resulting effect.
3. **Bounded causal interpretation.** Reverse, random, population, mismatched-donor, shuffled-ridge, and prefix controls separate directional boundary-state control from generic perturbation and exact query-specific restoration.

The resulting evidence characterizes a causal control region within the evaluated executions; it does not posit a universal refusal direction or an unconditional deployment defense.

2 Related Work

Jailbreak construction and behavioral evaluation. Instruction tuning (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022) establish the refusal–compliance setting formalized by Wei et al. (Wei et al., 2023). Gradient-based suffix optimization (Zou et al., 2023b), GPTFuzzer (Yu et al.,

2023), AutoDAN (Liu et al., 2024), PAIR (Chao et al., 2025), and adaptive attacks (Andriushchenko et al., 2025) expose this gap under different access models. Functional Homotopy (Wang et al., 2025d), H4RM3L (Doumbouya et al., 2025), AutoDAN-Turbo (Liu et al., 2025), One Model Transfer to All (Li et al., 2025a), Adversarial Deja Vu (Dabas et al., 2026), Auto-RT (Liu et al., 2026), SlotGCG (Jeong et al., 2026), and self-jailbreaking (Yong & Bach, 2026) extend attacks across optimization, transfer, positional, and reasoning settings. HarmBench (Mazeika et al., 2024), JailbreakBench (Chao et al., 2024), and WildTeaming (Jiang et al., 2024a) standardize behaviors; StrongREJECT (Souly et al., 2024), SorryBench (Xie et al., 2025), and GuidedBench (Huang et al., 2026) improve judging; XSTest (Röttger et al., 2024), OR-Bench (Cui et al., 2024), and SafeDialBench (Cao et al., 2026) expose utility and multi-turn effects. These resources make outcomes comparable, but do not localize the hidden transition that redirects a complete generation.

Refusal representations and boundary mechanisms. Representation Engineering (Zou et al., 2023a) and CAA (Rimsky et al., 2024) show that safety features can be read or shifted in residual space. A single refusal direction (Arditi et al., 2024), concept activation vectors (Xu et al., 2024), and mechanistic safety-fine-tuning analyses (Jain et al., 2024) reveal structured but non-identical signals; refusal geometry (Wollschläger et al., 2025), COSMIC (Siu et al., 2025), and multilingual direction analysis (Wang et al., 2025c) test transfer. RepIt (Siu et al., 2026), AlphaSteer (Sheng et al., 2026), refusal-feature adversarial training (Yu et al., 2025), and vector ablation for false refusal (Wang et al., 2025b) turn these signals into interventions, while shared representations explain cross-model transfer (Angell et al., 2026). Template-region analysis (Leong et al., 2025) and token-depth experiments (Qi et al., 2025) place safety computation near the prompt-to-generation transition. Yet the common evidence is decodability or steerability, not a prompt-side state mediating complete responses.

Runtime safety and causal intervention. Conditional activation steering (Lee et al., 2025), dynamic vectors (Wang et al., 2025a), controllable alignment (Zhang et al., 2025), Jailbreak Antidote (Shen et al., 2025), SmoothLLM (Robey et al., 2023), RobustKV (Jiang et al., 2024b), and SALORA (Li et al., 2025b) optimize safety–utility trade-offs at inference or via parameter updates; ProSafePrune (Chen et al., 2026), Discern (Lu et al., 2026), Any-depth Alignment (Zhang et al., 2026), and ASGuard (Park et al., 2026) target over-refusal and runtime protection. These methods address deployment, not matched mechanistic counterfactuals. ROME-style causal tracing (Meng et al., 2022), Patchscopes (Ghandeharioun et al., 2024), activation-patching best practices (Zhang & Nanda, 2024), and subspace-patching analyses (Makelov et al., 2024) provide intervention tools and cautions, but do not connect a jailbreak-specific state change to a full response trajectory.

Positioning our analysis. Our study combines these threads with matched direct/wrapped renderings, query-group-disjoint donor prediction, one-state replacement at the assistant-generation boundary, complete-generation scoring, and prefix mediation. This design separates directional control from perturbation magnitude, exact-query matching, and first-token correlation, enabling tests of where the jailbreak transition is expressed and how it propagates through autoregressive generation.

3 Causal Analysis at the Assistant-Generation Boundary

3.1 Problem formulation and causal estimand

Let f be a frozen decoder-only chat model with residual-stream state $h_{\ell,p}(x) \in \mathbb{R}^d$ at layer ℓ and prompt position p. For harmful query q_i, x_i^D is its direct chat rendering and x_{ij}^W is the same query embedded in wrapper family j; the paired task is therefore held fixed.

We call the final prompt-side position immediately preceding assistant decoding the *assistant-generation boundary*. For candidate (ℓ, p), the paired displacement is

$$\Delta_{ij}^{\ell,p} = h_{\ell,p}(x_i^D) - h_{\ell,p}(x_{ij}^W). \tag{1}$$

The direct state is a query-matched empirical counterfactual, not a normative target. We use "causal" in the intervention sense: the prompt, model, decoding rule, and all other activations remain fixed

while one state changes. Our estimand is the paired change in generation-level outcome Y when a wrapped state is replaced by a predicted direct-like state:

$$\tau = \mathbb{E}_{i,j}\left[Y\Big(f(x_{ij}^W; \mathrm{do}(h_{\ell^\star,p^\star} \leftarrow \widehat{d}_{ij}))\Big) - Y\big(f(x_{ij}^W)\big)\right], \tag{2}$$

where $Y = 1$ denotes substantive harmful compliance and $\widehat{d}_{ij}$ is estimated without the test query's direct run. Equation 2 is therefore an execution-conditional state effect, not a claim of universal necessity or a deployable defense.

These definitions yield three tests: sensitivity should concentrate near the boundary rather than arbitrary content positions; a direct-like move should reduce Y while its reverse increases it relative to a norm-matched random move; and a short generated prefix should carry a substantial, model-dependent fraction of the full effect. The evidence can localize a band, but not a unique causal layer.

3.2 Localizing an intervention-sensitive region

Candidate positions include harmful-content tokens, wrapper tokens, chat-template markers, and the final prompt token. First, a validation-only refusal-probe sweep selects one primary coordinate $(\ell^\star, p^\star)$ per model. Separately, a held-out generation-level grid evaluates the broader pattern across layer–position and content/wrapper controls; it is descriptive and cannot change the frozen primary choice or use final-set labels. At the selected coordinate, write

$$w_{ij} = h_{\ell^\star,p^\star}(x_{ij}^W), \qquad d_i = h_{\ell^\star,p^\star}(x_i^D). \tag{3}$$

The probe establishes decodability only; the held-out generation intervention supplies the causal test. Full grid sizes, generation budgets, and uncertainty calculations are given in Appendix A.

3.3 Predicting and patching a direct-like state

An oracle could insert d_i into the wrapped run, but would require the test query's direct execution. We instead learn a direct-like counterfactual from training pairs. Stack centered wrapped states into X_c and direct targets into D_c, and fit the multi-output ridge map

$$W^\star = \arg\min_W \|X_c W - D_c\|_F^2 + \alpha\|W\|_F^2, \qquad \widehat{d}(w) = \bar{d} + (w - \bar{w})W^\star, \tag{4}$$

with fixed $\alpha = 10$. Each training query supplies its wrapped states and one direct target; query-group-disjoint splits prevent the intent or a rephrasing from crossing fitting, validation, and test. During a test wrapped pass, we cache w, replace only $h_{\ell^\star,p^\star}$ with $\widehat{d}(w)$, and resume ordinary greedy decoding; no weights, masks, or other states are edited.

Ridge is a transparent donor-construction device, not the claimed mechanism; the causal claim concerns the resulting state intervention rather than predictor sophistication.

3.4 Intervention family and geometric controls

The primary *predicted* intervention inserts $\widehat{d}(w)$. The *direct* control inserts the same query's d_i as an oracle reference; *reverse* uses $w - (\widehat{d}(w) - w)$; and *random* adds an isotropic displacement with the same norm. Specificity controls use a different-query direct state, global means, and a shuffled ridge map. Together they separate directional restoration from exact query matching, generic perturbation, and state normalization.

To expose the geometry, define the shared displacement

$$v_g = \mathbb{E}_{i,j}[d_i - w_{ij}]. \tag{5}$$

We separately intervene with the *global* donor $w + v_g$, the adaptive residual $\widehat{d}(w) - (w + v_g)$, and their scaled combination

$$z(\lambda_g, \lambda_r) = w + \lambda_g v_g + \lambda_r\big[\widehat{d}(w) - (w + v_g)\big]. \tag{6}$$

Dose, singular-spectrum, projection-removal, and benign-population analyses use these definitions to distinguish a universal axis, prompt-adaptive counterfactual geometry, and generic manifold re-entry; their implementations and results are reported in Appendix C.

3.5 Evaluation logic and leakage control

The primary endpoint is attack success rate (ASR) on fresh, length-matched generations, judged by substantive harmful compliance rather than lexical refusal; probes and first-token statistics are diagnostics. For prefix mediation, we generate baseline and patched prefixes, force each length $k \in \{1, 2, 4, 8\}$, and resume decoding. Let Y_{BB}, Y_{BP}, Y_{PB}, and Y_{PP} denote the four state–prefix combinations. We report

$$\begin{aligned} T(k) &= \mathrm{ASR}_{BB} - \mathrm{ASR}_{PP}, \quad P(k) = \mathrm{ASR}_{BB} - \mathrm{ASR}_{BP}, \\ R(k) &= \mathrm{ASR}_{BP} - \mathrm{ASR}_{PP}, \quad M(k) = \frac{P(k)}{T(k)}, \end{aligned} \tag{7}$$

where $M(k)$ is the recovered fraction of the total effect. This decomposition tests early-trajectory mediation without treating one token as a unique mediator.

Layers, positions, hyperparameters, and thresholds are fixed on training/validation groups; final and external behaviors are disjoint from fitting. Paired McNemar tests and query-group bootstrap intervals retain within-group wrapper correlation, and all controls use matched decoding. The 2×2 token clamp provides an additional prefix-mediation check. Appendix A gives the complete protocol.

4 Experiments

We organize the experiments around six questions: where sensitivity concentrates; whether one-state intervention changes complete generations; how much of that change appears in the initial prefix; whether non-matched donors suffice; what geometry the intervention uses; and whether the effect transfers beyond the fitting distribution.

Setup. We evaluate Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Gemma-2-9B-it on 933 harmful queries from AdvBench, HarmBench, and JailbreakBench (JBB), each paired with seven wrapper families. We use 600 groups for fitting, 133 for selection, and a 200-group held-out pool for final hidden-state controls; Table 1 uses its balanced 20-group slice (140 pairs/model). Choices are frozen before generation. The primary endpoint is judged ASR. A 504-output human audit obtains 98.21% agreement with the primary judge ($\kappa = .964$); Appendix A reports judging and uncertainty details.

Evaluation packages. Table 1 uses the primary 256-token generation protocol. Behavioral localization and the factorial clamp use separately frozen 64-token mechanism protocols. Absolute ASR values should therefore be compared only within a matched panel or table, where generation and judging conditions are identical.

4.1 Where does generation-level sensitivity concentrate?

A validation-only probe sweep selects one coordinate per model. We then test the broader spatial pattern with a held-out generation-level grid across 96 layer-by-position coordinates. The grid uses 20 held-out query groups, seven wrappers, 140 query–family pairs per model, greedy 64-token generations, independently fitted donor maps, and group-bootstrap intervals. Each cell reports paired $\Delta\mathrm{ASR} = \mathrm{ASR}_{\mathrm{patched}} - \mathrm{ASR}_{\mathrm{baseline}}$; negative values indicate fewer successful jailbreaks. The grid is descriptive. It does not select the primary coordinate or tune hyperparameters.

Behavioral sensitivity concentrates near B/B-1 in late layers, while content and wrapper controls remain close to zero. Qwen peaks at L16/B, Llama at L20/B, and Gemma at L20/B-1. Llama L12/B-2 is also substantial, reinforcing a structured neighborhood. The evidence supports a model-dependent band, not a universal boundary token or layer.

4.2 Does restoring the boundary state change complete generations?

Having localized behavioral sensitivity to a late boundary band, we test whether the frozen primary coordinate controls complete generations under a fixed prompt and decoding rule. Table 1 reports 140 held-out query–family pairs per model. The predicted donor lowers ASR from 70.7% to 56.4%

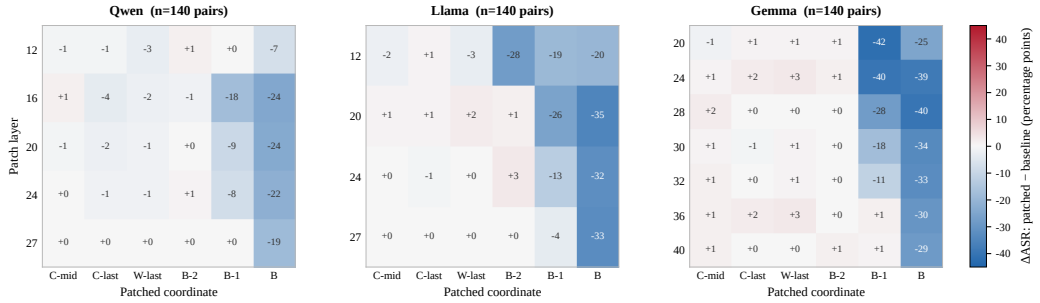

Figure 2: **Generation-level behavioral localization across 96 coordinates.** Each cell reports paired $\Delta\text{ASR} = \text{ASR}_{\text{patched}} - \text{ASR}_{\text{baseline}}$ on 140 held-out query–family pairs per model; negative values indicate reduced jailbreak success. Columns compare content-middle (C-mid), content-last (C-last), wrapper-last (W-last), and three late boundary positions (B-2, B-1, B). Descriptive peaks occur at Qwen L16/B (-24.3 points), Llama L20/B (-35.0 points), and Gemma L20/B-1 (-42.1 points). Late B/B-1 averages range from -13.1 points for Qwen to -26.1 points for Gemma; control averages stay from -0.8 to $+1.6$ points. The exact peak differs by model, supporting a model-dependent late boundary band.

Table 1: **Single-state intervention changes complete generations.** Judged ASR (%) on 140 held-out query–family pairs per model. PREDICTED inserts the ridge-predicted direct-like state; DIRECT inserts the same-query oracle state; REVERSE moves by the predicted displacement in the opposite direction; and RANDOM applies a norm-matched isotropic perturbation. Fix/Reg. counts baseline jailbreaks corrected by the predicted donor and baseline-safe generations made unsafe.

Model	Base	**Predicted**	Direct	Reverse	Random	Fix/Reg.
Qwen	70.7	56.4	55.7	81.4	71.4	23/3
Llama	63.6	29.3	30.7	69.3	61.4	50/2
Gemma	60.7	23.6	22.9	67.9	64.3	53/1

for Qwen (95% query-cluster bootstrap CI $[6.4, 21.4]$, $p < 10^{-4}$), from 63.6% to 29.3% for Llama ($[26.4, 42.1]$, $p < 10^{-12}$), and from 60.7% to 23.6% for Gemma ($[28.6, 45.0]$, $p < 10^{-12}$).

Three controls constrain the interpretation. The same-query oracle donor produces nearly the same ASR as the predicted donor, and paired tests do not detect a difference ($p \geq .77$). Reverse interventions increase ASR for every model, showing that direction matters. Norm-matched random perturbations remain near baseline on this primary package, so perturbation magnitude alone does not explain the learned direction here. Fix/Reg. counts show that corrections substantially outnumber regressions, although the intervention is not deterministic. Frozen-coordinate extensions preserve the main direction, with a larger random shift on the extended Qwen slice (Appendix E).

4.3 HOW MUCH OF THE EFFECT IS CARRIED BY A SHORT GENERATED PREFIX?

The patch acts before any assistant token is emitted. Refusal-start mass rises from 0.198 to 0.939 for Qwen, from 0.309 to 0.947 for Llama, and from 0.248 to 0.974 for Gemma. First-token entropy decreases from 0.417 to 0.168, from 0.436 to 0.136, and from 0.502 to 0.121. The top token changes on 80.7%, 65.7%, and 75.0% of examples. These diagnostics show an early policy shift. They do not by themselves establish mediation.

We force prefixes with $k \in \{1, 2, 4, 8\}$ in the pre-specified 2×2 design from Eq. 7. Let Y_{BB}, Y_{BP}, Y_{PB}, and Y_{PP} denote baseline-state/baseline-prefix, baseline-state/patched-prefix, patched-state/baseline-prefix, and patched-state/patched-prefix outcomes. Figure 3 reports the total effect $T(k)$, prefix component $P(k)$, and residual state-after-prefix effect $R(k)$.

Across the tested lengths, the prefix-mediated fraction ranges from 84.8% to 93.9% for Qwen, from 83.9% to 87.1% for Llama, and from 89.1% to 100.0% for Gemma. The residual effect is only from 1.4 to 3.6 points for Qwen, from 2.9 to 3.6 points for Llama, and from 0.0 to 3.6 points for Gemma.

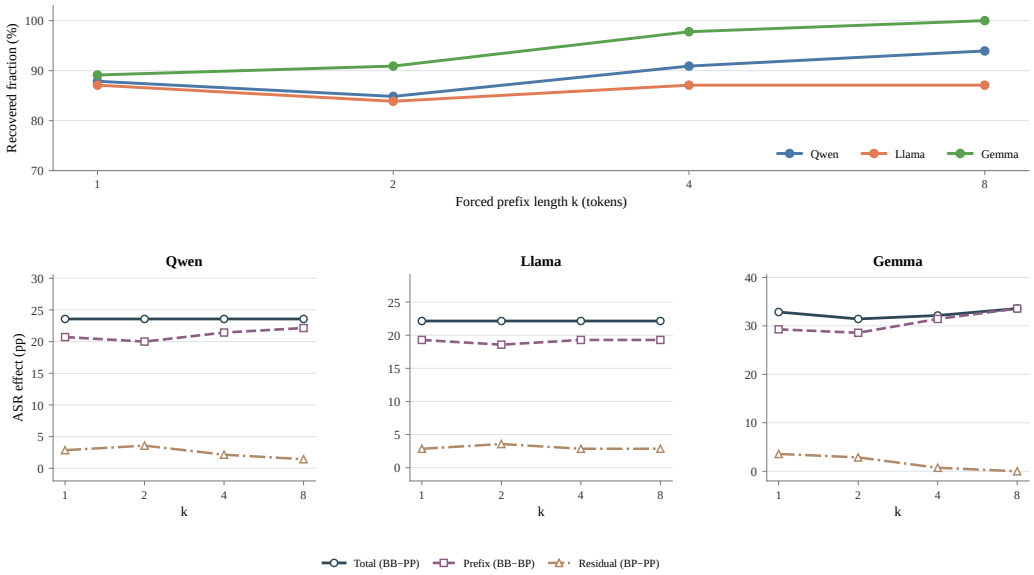

Figure 3: **Short-prefix mediation of the full-generation intervention effect.** The upper panel shows the recovered fraction $M(k) = P(k)/T(k)$ as the forced prefix length grows from 1 to 8 tokens. The lower panels report total effects $T(k) = \mathrm{ASR}_{BB} - \mathrm{ASR}_{PP}$, prefix effects $P(k) = \mathrm{ASR}_{BB} - \mathrm{ASR}_{BP}$, and residual effects $R(k) = \mathrm{ASR}_{BP} - \mathrm{ASR}_{PP}$. Prefixes recover most of the total effect across all three models; residual effects remain small and model dependent.

Most downstream behavior is therefore transmitted through a short initial generated prefix. This moves the evidence beyond first-token correlation: the state intervention changes an early trajectory that carries most of the eventual generation-level effect. It does not imply that one token is the unique mediator or that later dynamics are causally irrelevant. Absolute ASRs for all four forced conditions appear in Appendix B; harmful-versus-benign counterfactuals provide complementary evidence in Appendix C.

4.4 Can simpler state-restoration explanations account for the effect?

A direct-like patch could return the wrapped state to a generic in-distribution region. Figure 4 shows that the norm-matched benign-population anchor nearly matches the predicted donor for Llama and Gemma but is weaker for Qwen, whereas projection removal is small or changes sign. The boundary is causally effective, yet its geometry need not equal exact query-specific restoration; the underlying counts are in Appendix D.

4.5 How specific must the boundary-state restoration be?

We next evaluate the stronger specificity question directly. The frozen package reuses the same 20 query groups and seven wrapper families for all three models, generates 256 tokens greedily, and judges all nine variants in one independently recorded package per model. Figure 5 reports the within-package ASRs; the full point-estimate table is given in Appendix D. Its absolute baselines are not pooled with Table 1 because the judge calls were made in a separate batch.

The predicted donor reduces ASR relative to its package baseline for all three models. The mean-displacement donor remains close for Qwen but is weaker for Llama and Gemma; the other non-matched donors frequently overlap the predicted estimate in Figure 5. Predicted-versus-direct, mismatched, mean-direct, and shuffled-ridge McNemar tests provide no reliable evidence that exact query matching is necessary or consistently superior ($p \geq .0625$ across models and controls), whereas predicted-versus-reverse and predicted-versus-random comparisons are directional for all three models (all $p \leq 1.2 \times 10^{-4}$). These controls support transfer of a causal control region rather than a particular donor estimator; the full point estimates are in Appendix D.

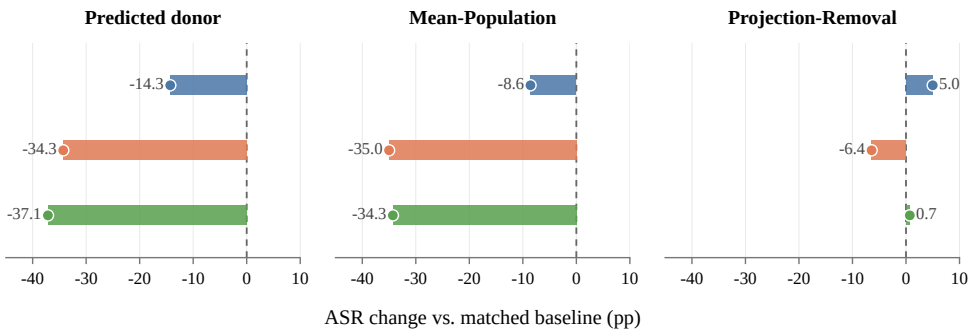

Figure 4: **Alternative explanations for boundary-state repair.** Horizontal bars show the change in judged ASR relative to the matched baseline; negative values indicate fewer successful jailbreaks. The predicted donor lowers ASR for all three models. The norm-matched benign-population anchor reproduces much of the Llama and Gemma reduction but less of the Qwen reduction, whereas projection removal is small and changes sign across models.

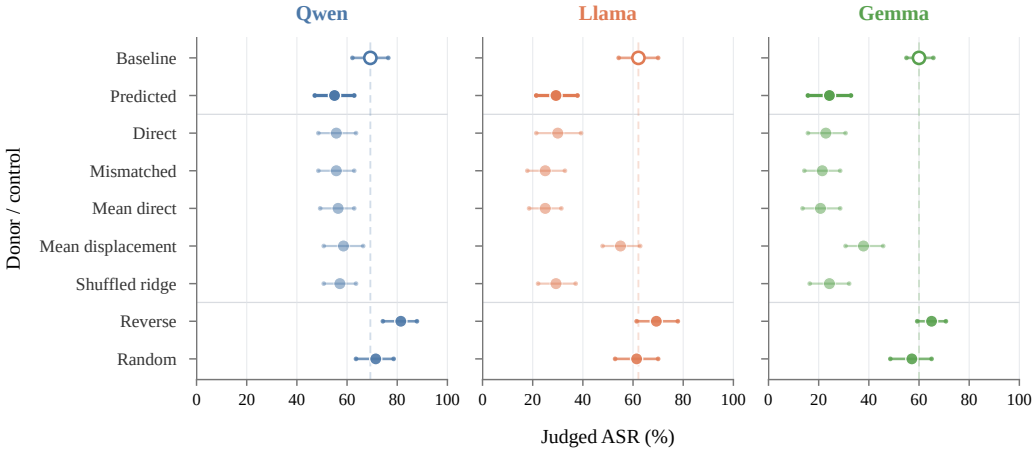

Figure 5: **Generation-level donor-specificity controls.** Each model panel reports judged ASR for nine variants on the same 140 held-out query–family pairs, using a separate fixed judge batch. Points are ASR estimates and horizontal intervals are 95% query-cluster bootstrap intervals; the hollow point and dashed vertical line mark the model-specific baseline. Same-query direct, mismatched-direct, mean-direct, and shuffled-ridge donors often overlap the matched predicted donor. The mean-displacement donor is less effective for Llama and Gemma, while reverse and norm-matched random controls remain near or above baseline.

4.6 What geometric structure does the intervention use?

The intervention contains a recurring shared component but is not uniformly described by one direction. Figure 6 shows that the global component is behaviorally active, while the complete donor is stronger and residual effects vary across models.

The fitted ridge maps have broad spectra. Without rank-constrained generation interventions, singular values do not identify causal dimensionality; the remaining representation diagnostics are in Appendix C.

Scope under distribution shift. With the coordinate and ridge map frozen, the intervention transfers beyond the fitting families, but not uniformly. HarmBench and WildJailbreak both show positive aggregate changes, while family-level slices include neutral and adverse cases. These results support partial transfer of a boundary-related state family, not a universal donor estimator; Appendix F gives full results.

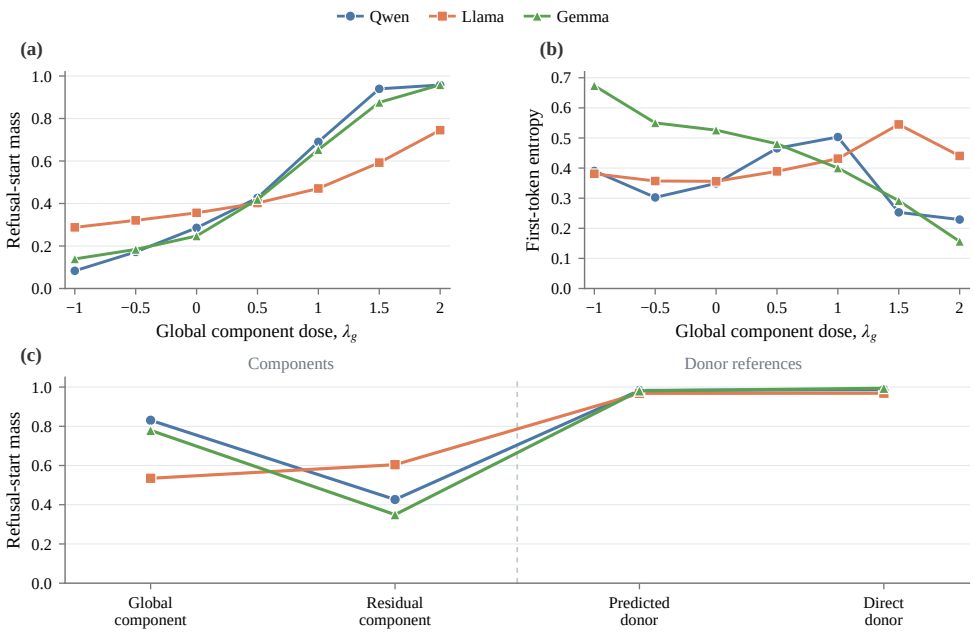

Figure 6: **Dose response and decomposition of the boundary intervention.** (a) Refusal-start mass and (b) first-token entropy as the dose λ_g of the shared global component varies. Refusal-start mass increases monotonically for all three models, whereas entropy changes are model dependent. (c) Refusal-start mass under the isolated global and residual components, followed by the full predicted and same-query direct donor references. The global component is behaviorally active, but the complete donors are stronger and more consistent; the residual component alone varies substantially across models.

5 Discussion and Limitations

Across models, generation-level sensitivity concentrates in a late assistant-generation boundary band, and one-state replacement changes complete jailbreak outcomes. A short initial prefix carries most of the downstream effect; reverse controls establish directionality, while specificity controls find no reliable need for exact query matching. The band is model dependent, not a uniquely identified layer or universal bottleneck. The boundary is therefore best interpreted as a control region linking prompt-side computation to the early autoregressive trajectory, rather than as a new steering algorithm.

The controls bound this interpretation. A benign-population anchor nearly matches the adaptive donor for Llama and Gemma, whereas mismatched direct, mean-direct, and shuffled-ridge donors are not reliably weaker than the matched predictor. A shared displacement is active, but projection-removal and residual effects vary by model; broad ridge spectra do not establish causal dimensionality.

The study covers three open-weight models, greedy decoding, seven wrapper families, and model-based judging. Human auditing supports the primary judge but not every output, absolute ASR varies across frozen packages, and transfer is heterogeneous. Unconditional patching increases benign refusal for Qwen and Llama and produces a Gemma OR-Bench safety regression. The claim is consequently limited to intervention-based mechanistic evidence under the evaluated models and decoding rule.

6 Conclusion

Jailbreak behavior is causally controllable from a model-dependent late assistant-generation boundary region. One-state replacement changes complete responses, with most of the effect carried by a short initial prefix. The evidence supports directional control of a refusal-policy region, not a universal vector, unique bottleneck, or deployable defense.

REFERENCES

Maksym Andriushchenko, Nicolas Flammarion, et al. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *International Conference on Learning Representations*, volume 2025, pp. 40116–40143, 2025.

Rico Angell, Jannik Brinkmann, and He He. Jailbreak transferability emerges from shared representations. In *International Conference on Learning Representations*, volume 2026, pp. 135328–135363, 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Hongye Cao, Sijia Jing, Yanming Wang, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Jiaheng Liu, Boyan Wang, et al. Safedialbench: A fine-grained safety evaluation benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. In *International Conference on Learning Representations*, volume 2026, pp. 125737–125785, 2026.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

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.

Zijun Chen, Wenbo Hu, Ya Li, Lei Miao, Guoping Hu, and Richang Hong. Prosafeprune: Projected safety pruning for mitigating over-refusal in llms. In *International Conference on Learning Representations*, volume 2026, pp. 18673–18695, 2026.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Mahavir Dabas, Tran Huynh, Nikhil Billa, Jiachen Tianhao Wang, Peng Gao, Charith Peris, Yao Ma, Rahul Gupta, Ming Jin, Prateek Mittal, et al. Adversarial déjà vu: Jailbreak dictionary learning for stronger generalization to unseen attacks. In *International Conference on Learning Representations*, volume 2026, pp. 104731–104769, 2026.

Moussa Koulako Bala Doumbouya, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky, and Christopher Manning. h4rm3l: A language for composable jailbreak attack synthesis. In *International Conference on Learning Representations*, volume 2025, pp. 57253–57286, 2025.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

Ruixuan Huang, Xunguang Wang, Zongjie Li, Daoyuan Wu, and Shuai Wang. Guidedbench: Measuring and mitigating the evaluation discrepancies of in-the-wild llm jailbreak methods. In *International Conference on Learning Representations*, volume 2026, pp. 110014–110043, 2026.

Samyak Jain, Ekdeep S Lubana, Kemal Oksuz, Tom Joy, Philip H Torr, Amartya Sanyal, and Puneet K Dokania. What makes and breaks safety fine-tuning? a mechanistic study. *Advances in Neural Information Processing Systems*, 37:93406–93478, 2024.

Seungwon Jeong, Jiwoo Jeong, Hyeonjin Kim, Yunseok Lee, and Woojin Lee. Slotgcg: Exploiting the positional vulnerability in llms for jailbreak attacks. *arXiv preprint arXiv:2606.05609*, 2026.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165, 2024a.

Tanqiu Jiang, Zian Wang, Jiacheng Liang, Changjiang Li, Yuhui Wang, and Ting Wang. Robustkv: Defending large language models against jailbreak attacks via kv eviction. *arXiv preprint arXiv:2410.19937*, 2024b.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pp. 90960–90985, 2025.

Chak Tou Leong, Qingyu Yin, Jian Wang, and Wenjie Li. Why safeguarded ships run aground? aligned large language models' safety mechanisms tend to be anchored in the template region. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15212–15229, 2025.

Linbao Li, Yannan Liu, Daojing He, and Yu Li. One model transfer to all: On robust jailbreak prompts generation against llms. In *International Conference on Learning Representations*, volume 2025, pp. 6359–6375, 2025a.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Salora: Safety-alignment preserved low-rank adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 90827–90843, 2025b.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, volume 2024, pp. 56174–56194, 2024.

Xiaogeng Liu, Peiran Li, G Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. In *International Conference on Learning Representations*, volume 2025, pp. 10313–10360, 2025.

Yanjiang Liu, Shuheng Zhou, Yaojie Lu, Huijia Zhu, Weiqiang Wang, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. Auto-rt: Automatic jailbreak strategy exploration for red-teaming large language models. In *International Conference on Learning Representations*, volume 2026, pp. 69406–69453, 2026.

Yuxiao Lu, Lin Xu, Wenjun Li, Jie Shi, et al. Discern truth from falsehood: reducing over-refusal via contrastive refinement. In *International Conference on Learning Representations*, volume 2026, pp. 41917–41936, 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Yein Park, Jungwoo Park, and Jaewoo Kang. Asguard: Activation-scaling guard to mitigate targeted jailbreaking attack. In *International Conference on Learning Representations*, volume 2026, pp. 152313–152341, 2026.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pp. 54911–54941, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

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, 2024.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 21452–21479, 2025.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations*, volume 2026, pp. 25861–25904, 2026.

Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. Cosmic: Generalized refusal direction identification in llm activations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25534–25553, 2025.

Vincent Siu, Nathan Henry, Nicholas Crispino, Yang Liu, Dawn Song, and Chenguang Wang. Repit: Steering language models with concept-specific refusal vectors. In *International Conference on Learning Representations*, volume 2026, pp. 109807–109851, 2026.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In *International Conference on Learning Representations*, volume 2025, pp. 79334–79351, 2025a.

Xinpeng Wang, Chengzhi Martin Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *International Conference on Learning Representations*, volume 2025, pp. 33824–33843, 2025b.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. *Advances in Neural Information Processing Systems*, 38:32380–32423, 2025c.

Zi Wang, Divyam Anshumaan, Ashish Hooda, Yudong Chen, and Somesh Jha. Functional homotopy: Smoothing discrete optimization via continuous parameters for llm jailbreak attacks. In *International Conference on Learning Representations*, volume 2025, pp. 38125–38144, 2025d.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, volume 2025, pp. 59937–59973, 2025.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector. *Advances in Neural Information Processing Systems*, 37:116743–116782, 2024.

Zheng Xin Yong and Stephen Bach. Self-jailbreaking: Language models can reason themselves out of safety alignment after benign reasoning training. In *International Conference on Learning Representations*, volume 2026, pp. 96061–96086, 2026.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust llm safeguarding via refusal feature adversarial training. In *International Conference on Learning Representations*, volume 2025, pp. 5254–5277, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Jiawei Zhang, Andrew Estornell, David Baek, Bo Li, and Xiaojun Xu. Any-depth alignment: Unlocking innate safety alignment of llms to any-depth. In *International Conference on Learning Representations*, volume 2026, pp. 25905–25939, 2026.

Jingyu Zhang, Ahmed Elgohary Ghoneim, Ahmed Magooda, Daniel Khashabi, and Ben Van Durme. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. In *International Conference on Learning Representations*, volume 2025, pp. 68007–68038, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

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*, 2023b.

A Experimental Protocol

Data and splits. The canonical pool contains 933 harmful queries aggregated from AdvBench, HarmBench, and JailbreakBench (JBB), each paired with seven wrapper families. Query groups 0–599, 600–732, and 733–932 are used for fitting, selection, and the held-out final pool, respectively. Hidden-state controls use all 200 final groups (1,400 wrapper rows per model); the balanced generation experiment uses 733–752 (20 groups; 140 query–family pairs per model). Frozen-coordinate extensions use Qwen 753–792 (280 pairs) and Llama/Gemma 753–772 (140 pairs each), without revising any intervention choice.

Table 2: Frozen intervention coordinates and generation settings. Positions refer to the model's rendered chat template.

Model	Layer	Position	Max tokens	Decoding
Qwen2.5-7B-Instruct	20	final newline	256	greedy
Llama-3.1-8B-Instruct	24	final newline	256	greedy
Gemma-2-9B-it	30	final newline	256	greedy

Judging and uncertainty. The primary judge marks substantive harmful compliance, safe content, and explicit refusal. Safe and refusal labels overlap because refusal is a subset of safe behavior. All paired variants for a comparison are submitted in the same evaluation package. We report exact two-sided McNemar tests for paired binary outcomes and 95% bootstrap intervals over query groups. A 504-output human audit covering 252 pairs found 98.21% agreement with the primary judge ($\kappa = .964$). Among complete-agreement pairs, the predicted patch produced 65 fixes, 4 regressions, 103 safe-safe outcomes, and 71 harmful-harmful outcomes.

Reproducibility choices. The ridge penalty is fixed at $\alpha = 10$. All representation statistics use held-out query groups. Generation variants within each comparison share token budgets and decoding parameters, and random controls match the intervention norm. External behaviors are normalized and checked for exact overlap with the fitting pool. Negative family and utility slices are reported rather than absorbed into a pooled score. The 140-pair prefix package uses groups 733–752, seven wrapper families, greedy decoding with 64 new tokens, prefix lengths $k \in \{1, 2, 4, 8\}$, and one judge endpoint for all forced conditions. Behavioral localization uses groups 753–772, 20 query groups, seven wrappers, 140 pairs per model, greedy 64-token generations, and independently fitted donor maps at 96 coordinates (Qwen 30, Llama 24, Gemma 42). Its deterministic judge is fixed across all cells. The donor-specificity package reuses groups 733–752 and all seven families, evaluates nine variants at 256 greedy tokens, uses a separate fixed judge batch per model, and is audited independently for duplicate keys, missing variants, judge errors, and query-cluster uncertainty.

B Prefix Mediation Diagnostics

The main text reports the mediation decomposition because it compares effects across prefix lengths. Figure 7 gives the absolute ASRs for the four forced conditions. The same state and prefix labels apply at each k: BB is baseline state with the baseline prefix, BP is baseline state with the patch prefix, PB is patched state with the baseline prefix, and PP is patched state with the patch prefix. The figure preserves the raw conditions for auditability; it does not change the primary generation endpoint.

An earlier single-token factorial slice produced a smaller Qwen token-only estimate. A separate independent 70-pair Qwen follow-up yielded ASR from 57.1% for free baseline decoding to 35.7% for free patched decoding, with 55.7% after forcing the baseline-selected token under the patched state and 38.6% after forcing the patch-selected token under the baseline state. The dedicated prefix-length package is the primary mediation result; the earlier slice is retained as a diagnostic rather than a separate mechanism claim.

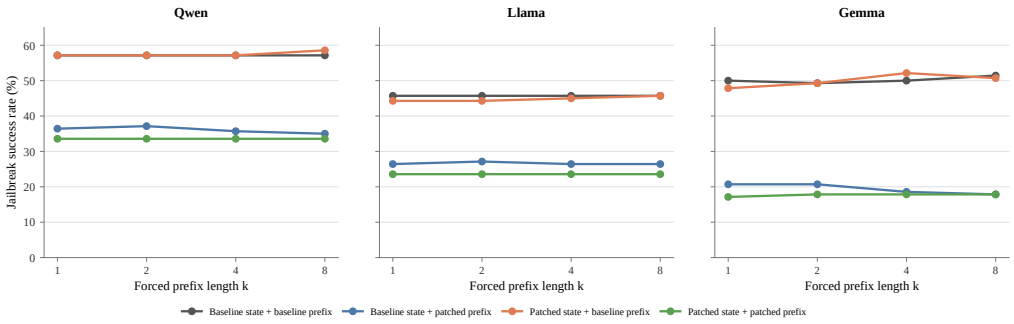

Figure 7: **Absolute ASR for the prefix-forcing conditions.** Each panel reports BB, BP, PB, and PP at forced prefix lengths from 1 to 8 tokens. The main text converts these conditions into total, prefix, and residual effects.

C Representation Evidence

On held-out baseline-jailbreak examples, the refusal probe separates the wrapped state from the high-refusal region and moves the predicted donor toward that region. Mean baseline/predicted refusal scores (R) are 0.235/0.975 for Qwen, 0.167/0.994 for Llama, and 0.304/0.850 for Gemma. A separate harmful-content probe (H) remains high: 0.997/0.981, 0.998/0.999, and 1.000/0.980. Thus the intervention restores refusal geometry without simply erasing detectable harmful content.

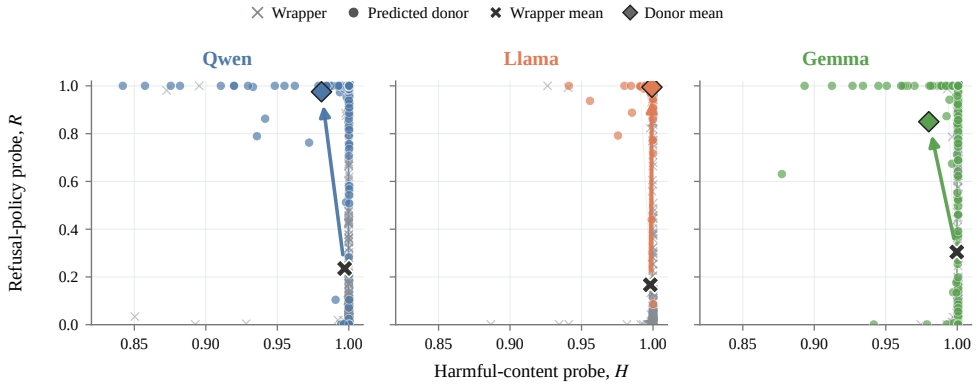

Figure 8: **Probe-space movement on held-out baseline-jailbreak examples.** Each panel plots the harmful-content probe H against the refusal-policy probe R. Gray crosses and model-colored circles denote wrapped states and their predicted donors; the black cross and colored diamond mark the corresponding means. Sparse connectors show paired examples, and the arrow shows the mean shift. Donors move sharply upward in R while remaining near the high-H region, indicating restored refusal geometry without wholesale removal of harmful-content information.

Figure 9 makes the model dependence explicit. Qwen and Llama exhibit large probe gains over broad portions of the cosine range, whereas Gemma does not show the same monotone alignment–gain relation. High cosine alignment with v_g is therefore neither a model-invariant predictor nor a complete account of refusal restoration.

Harmful versus benign wrappers. Direct-to-wrapped changes in refusal-probe score are -0.511 versus -0.426 for Qwen, -0.466 versus -0.285 for Llama, and -0.482 versus -0.435 for Gemma, comparing harmful and benign contents. The more policy-specific separation appears in refusal-start mass: harmful changes are -0.630, -0.589, and -0.697, whereas benign changes are -0.049, -0.192, and -0.322.

First-token entropy shows a complementary content dependence. Wrapping raises entropy for harmful prompts by 0.546, 0.485, and 0.523 for Qwen, Llama, and Gemma, respectively, but changes benign-

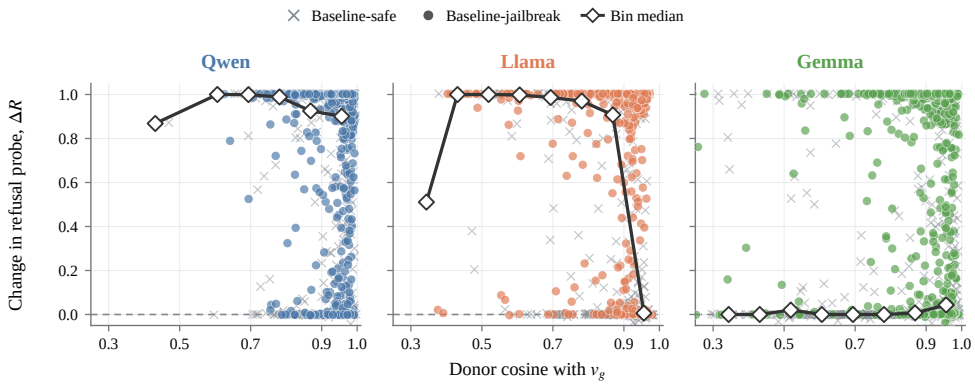

Figure 9: **Alignment of predicted-donor displacements with the shared direction.** The horizontal axis is the cosine between $\hat{d}(w) - w$ and the global mean displacement v_g; the vertical axis is the resulting change in refusal-probe score, ΔR. Gray crosses are baseline-safe examples and model-colored circles are baseline-jailbreak examples. Hollow diamonds connected by dark lines show cosine-bin medians. The alignment–gain relation differs sharply across models, so alignment with v_g alone does not explain the donor effect.

prompt entropy by -0.196, -0.144, and $+0.024$. Thus the wrapper-induced uncertainty increase is consistent across harmful contents but not across benign contents.

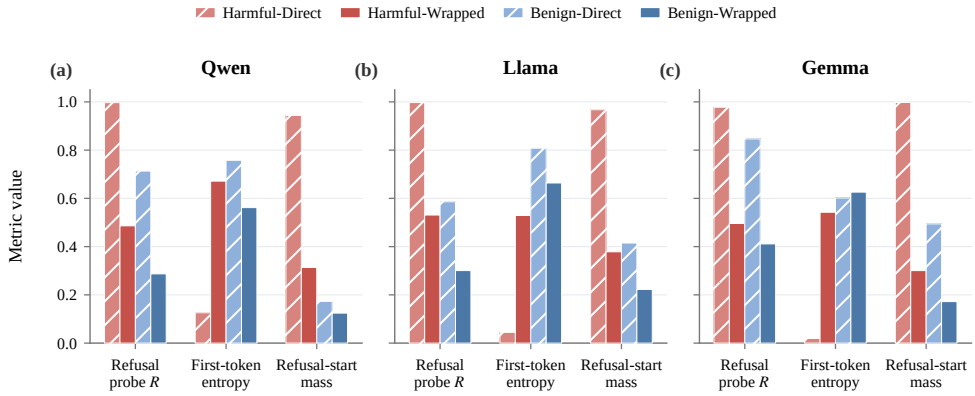

Figure 10: **Matched harmful and benign counterfactuals.** Panels (a)–(c) report refusal-probe score, first-token entropy, and refusal-start mass for Qwen, Llama, and Gemma. Red and blue bars denote harmful and benign content; hatched and solid bars denote direct and wrapped renderings. Wrapping lowers the refusal-probe score in both content classes, but the loss of refusal-start mass is substantially larger for harmful prompts. Harmful wrappers also increase first-token entropy for every model, whereas the benign entropy response is small or negative.

D Causal and Null Controls

Specificity package. The full point estimates underlying Figure 5 are listed below; the figure remains in the main text for visual comparison across donor variants.

State swaps and information retention. In a bidirectional 70-pair test, inserting the same-query direct state into wrapped contexts reduces ASR from 65.7% to 41.4% for Qwen and from 52.9% to 17.1% for Llama. Injecting the wrapped state into direct contexts raises ASR from 0% to 4.3% and 12.9%, respectively. A held-out TF-IDF decoder retrieves the canonical query in its top five at similar rates before and after patching (27.1% to 29.3% for Qwen, 42.1% to 48.6% for Llama, and 26.4% to 27.1% for Gemma), providing a coarse check against catastrophic content erasure.

Table 3: **Frozen specificity package.** Judged ASR (%) on 140 query–family pairs per model, with 256 greedy new tokens. Values are within-package point estimates; paired predicted-versus-control McNemar tests are reported in the text.

Model	Base	Pred.	Direct	Mismatch	Mean	MeanΔ	Shuffle	Reverse	Random
Qwen	69.3	55.0	55.7	55.7	56.4	58.6	57.1	81.4	71.4
Llama	62.1	29.3	30.0	25.0	25.0	55.0	29.3	69.3	61.4
Gemma	60.0	24.3	22.9	21.4	20.7	37.9	24.3	65.0	57.1

Population anchors and projection removal. Mean-Population replaces the boundary state with a norm-matched benign-state anchor. Projection-Removal subtracts the component along v_g while retaining the orthogonal wrapper residual. Table 4 gives the underlying ASR and paired transition counts for Figure 4. Generic manifold restoration explains much of the Llama and Gemma effect but less of the Qwen effect; removing the shared component is weak or inconsistent.

Table 4: Null-state and direction-removal controls on 140 held-out pairs. Delta is relative to the batch-matched baseline.

Model	Variant	ASR	Δ	Fix/Reg.
Qwen	Mean-Population	62.1	-8.6	17/5
	Projection-Removal	75.7	$+5.0$	1/8
Llama	Mean-Population	28.6	-35.0	50/1
	Projection-Removal	57.1	-6.4	13/4
Gemma	Mean-Population	26.4	-34.3	49/1
	Projection-Removal	61.4	$+0.7$	5/6

E Extended Generation Results

Frozen-coordinate extensions test whether the primary 20-group slice (733–752) is atypical. Qwen uses groups 753–792 (40 groups; 280 pairs), whereas Llama and Gemma use 753–772 (20 groups; 140 pairs each). Cross-model comparisons use the common 753–772 package; Qwen 773–792 is reported separately as a Qwen-only robustness slice. On 280 additional Qwen pairs, ASR changes from 73.2% to 24.3% (144 fixes, 7 regressions); direct, reverse, and random controls reach 22.9%, 78.2%, and 50.7%. Across 140 additional Llama pairs, baseline/predicted ASR is 56.4%/13.6% (62 fixes, 2 regressions in the extension), with direct, reverse, and random controls at 12.1%, 61.4%, and 53.6%. Across 140 additional Gemma pairs, baseline/predicted is 57.1%/24.3% (48 fixes, 2 regressions in the extension), and the corresponding controls are 24.3%, 65.7%, and 59.3%. The larger Qwen slice also shows a meaningful random shift, but the learned and direct interventions remain substantially stronger.

Separate 70-pair first-token follow-up. In an earlier frozen Qwen follow-up, the four conditions yield 57.1% ASR for free baseline decoding, 35.7% for free patched decoding, 55.7% for the patched state followed by the baseline-selected token, and 38.6% for the baseline state followed by the patch-selected token. This independent slice agrees with the 140-pair factorial result in Section 4.

F Transfer Across Families and Data Sources

Family-held-out fitting. Each attack family is excluded from donor fitting in turn. Within a separately judged family-held-out package on the same 140 prompts, baseline/predicted ASRs are 65.7%/56.4% for Qwen, 60.7%/25.0% for Llama, and 55.7%/32.1% for Gemma. These baselines differ from Table 1 because the judge calls are independently batched; comparisons are made within each package. Reverse controls reach 76.4%, 67.9%, and 60.7%. Transfer is positive in aggregate but not universal. Qwen PAIR remains at 100% and Qwen GCG at 30%, while Llama DeepInception remains at 60%; these are zero-change slices in Figure 11. Gemma GCG is also unchanged, and Gemma Jailbroken worsens by 5 percentage points.

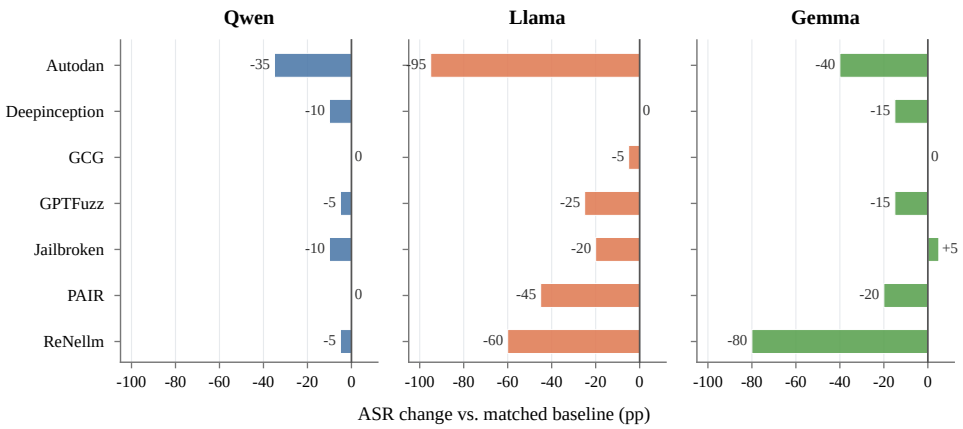

Figure 11: **Attack-family effects under frozen boundary interventions.** Each horizontal bar is predicted-donor ASR minus its matched baseline in percentage points. Negative values indicate reduced attack success, zero indicates no change, and positive values indicate a regression. Large reductions for several families coexist with neutral slices and a $+5$-point adverse Gemma Jailbroken slice, showing that transfer is heterogeneous rather than universal.

External behaviors and wrappers. The HarmBench experiment uses 160 exact-novel behaviors with a fixed AutoDAN Developer-Mode wrapper. The WildJailbreak experiment uses 200 unique paired vanilla and WildTeaming-generated prompts with zero normalized exact overlap with the 933-query pool. Figure 12 compares baseline ASR with the frozen ridge-predicted donor. On WildJailbreak, a mean-displacement donor (not plotted) reaches 50.5%, 46.0%, and 34.5% ASR, outperforming ridge for Gemma but not Llama; a global direct-state anchor is significant on all three. This ranking change is the clearest evidence that the causal state family transfers more reliably than any one estimator.

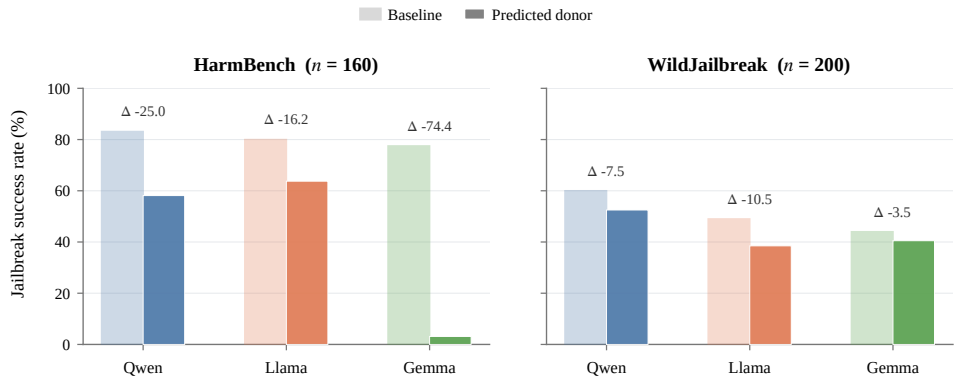

Figure 12: **External generalization of the frozen predicted donor.** Light and saturated bars show baseline and predicted-donor ASR, respectively; labels above each pair give the percentage-point change. HarmBench evaluates 160 exact-novel behaviors under a fixed AutoDAN Developer-Mode wrapper, whereas WildJailbreak evaluates 200 unique prompts that change both behavior and wrapper construction. The donor lowers ASR in every model–dataset pair, but the magnitude ranges from -74.4 points for Gemma on HarmBench to -3.5 points for Gemma on WildJailbreak.

G Utility, Failure Modes, and Judge Checks

Unconditional boundary repair can suppress benign answers as well as attacks. Figure 13 reports changes in refusal rate rather than raw counts. Qwen increases by 4, 20, and 13 percentage points on XSTest, OR-Bench, and FalseReject, respectively, while Llama increases by 15, 10, and 14

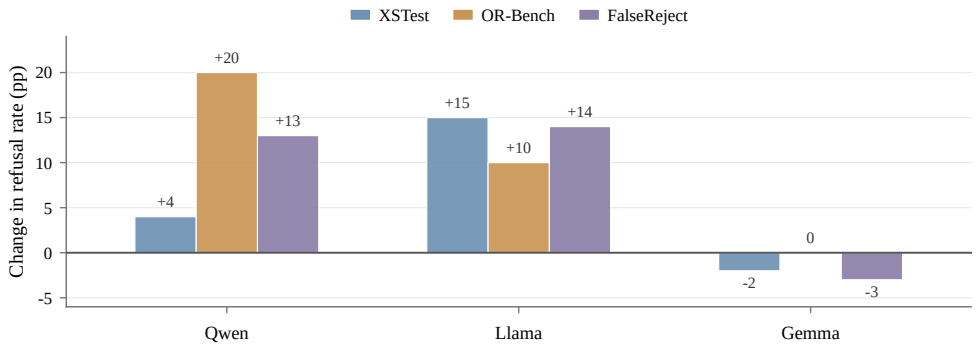

Figure 13: **Collateral behavior under unconditional intervention.** Bars show the change in refusal rate, in percentage points, relative to each model–dataset baseline; positive values indicate increased refusal and negative values indicate reduced refusal. Qwen and Llama become more refusing on all three benign benchmarks, while Gemma's refusal rate is flat or slightly lower. Refusal rate is not a complete safety metric: the zero Gemma OR-Bench bar coexists with a shift from safe answers to harmful compliance described in the text. Improvements on adversarial prompts therefore do not imply a favorable safety–utility operating point.

points. Gemma instead changes by -2, 0, and -3 points. The zero refusal-rate change for Gemma on OR-Bench does not imply preserved behavior: safe answers fall from 43 to 37 while harmful compliance rises from 7 to 13. This is a safety regression that refusal rate alone does not expose.

The primary judge was calibrated on 300 JailbreakBench outputs with three repeated calls. Complete rows achieved 93.0–94.4% agreement with human-majority labels ($\kappa = 0.845$–0.877); repeat agreement was $\kappa = 0.915$–0.946. A separate Llama-3.1-8B judge on 200 Qwen outputs assigned lower absolute ASR but preserved the intervention direction. The human audit reported in Appendix A further supports the paired comparisons. We nevertheless treat judged ASR as a measurement with uncertainty, not as ground truth.