RAS: MEASURING LLM SAFETY THROUGH REFUSAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated safety evaluation of open-weight large language models requires generating and judging responses for many related checkpoints. We introduce RAS (Refusal Alignment Score), a reference-relative measure of harmful-request refusal computed by the five-stage SafeVec procedure. SafeVec extracts layer-wise refusal directions from an aligned reference model, measures their alignment in related targets, and calibrates a 0–100 score using behavioral attack success rates (ASR) from calibration models. Subsequent target scoring uses input forward passes alone. Across 18 held-out variants from the Llama, Gemma, and Qwen families, RAS correlates with $100(1 - \mathrm{ASR})$ (Pearson 0.934, Spearman 0.851) and achieves 83.0% pairwise ordering agreement. Across five Llama prompt configurations, reference directions have mean cosine similarity 0.939. With calibration ASR already available, setup-inclusive evaluation of five targets is 3.2–$80.5\times$ faster than the specified behavioral protocol. These results support refusal alignment as an efficient signal for screening related checkpoints and prioritizing behavioral audits.

1 INTRODUCTION

Open-weight language models evolve through instruction tuning, specialization, model merging, and other post-training changes. Each resulting checkpoint brings a renewed evaluation burden: alignment established by human or AI feedback can change as the model is adapted (Ouyang et al., 2022; Bai et al., 2022). Fine-tuning can weaken safeguards even when harmful behavior is unintended (Qi et al., 2024), while tamper-resistance evaluations cover a broad range of weight- and representation-space modifications (Hossain et al., 2026). An evaluator auditing many related checkpoints therefore needs both detailed behavioral measurements and efficient signals for deciding which changes deserve closer inspection.

The standard approach elicits responses to harmful or adversarial requests and judges whether the responses comply. HarmBench, JailbreakBench, StrongREJECT, and SORRY-Bench provide important infrastructure for this process (Mazeika et al., 2024; Chao et al., 2024; Souly et al., 2024; Xie et al., 2025). Generating responses and running a judge for every checkpoint incurs recurring cost. The resulting attack success rate (ASR), the fraction of requests judged to elicit harmful compliance, also depends on the prompt distribution, decoding configuration, and judge. Recent audits document judge errors under changes in victim models and attack distributions (Schwinn et al., 2026; Gao, 2026). A complementary internal measurement could help evaluators track checkpoint changes while reserving detailed behavioral testing for consequential or uncertain cases.

Refusal-related representations provide a candidate signal. Refusal behavior can be influenced through directions in a model's residual stream (Arditi et al., 2024). Subsequent work identifies multidimensional refusal structure (Wollschläger et al., 2025) and distinguishes harmfulness recognition from the decision to refuse (Zhao et al., 2025). These findings motivate a specific measurement hypothesis: *among related model variants, reduced alignment with an aligned reference's refusal representation is informative about degradation in harmful-request refusal.* The relevant unit of comparison is the model checkpoint, evaluated under a common representation and prompt protocol.

We operationalize this hypothesis with RAS (Refusal Alignment Score), computed by the five-stage SafeVec procedure. For each family, SafeVec constructs refusal directions from an aligned reference, selects a layer window, and measures target alignment under direct harmful requests and

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.

jailbreak-framed requests. Calibration models with behavioral ASR labels establish a reference-relative 0–100 scale. Freezing the directions, window, and mapping makes this setup reusable: each subsequent target is scored using input forward passes, with target ASR reserved for validating the measurement.

The reference anchor makes checkpoint comparisons explicit, while calibration provides a common reporting range and adjusts the relative severity of alignment drops across families. Together, they allow RAS to compare variants within a family while keeping scores on a shared scale across families.

We make three contributions: (1) `RAS` and its five-stage `SafeVec` computation, with an analytical characterization of reference anchoring and ordering; (2) evaluation on 18 held-out variants, including family-level uncertainty and pairwise agreement; and (3) analyses of prompt transfer, score sensitivity, reference-derived calibration, and runtime amortization.

2 Related Work

Behavioral safety evaluation. Safety benchmarks address several measurement targets. RealToxicityPrompts studies toxic continuations, StereoSet and BBQ study bias, and TruthfulQA studies truthfulness (Gehman et al., 2020; Nadeem et al., 2021; Parrish et al., 2022; Lin et al., 2022); SafetyBench evaluates safety-related knowledge (Zhang et al., 2024). Our target is harmful-request refusal. Red teaming and jailbreak generation expose failures through model interaction (Ganguli et al., 2022; Perez et al., 2022; Chao et al., 2025; Zou et al., 2023b; Wei et al., 2023). Standardized refusal benchmarks and judges improve comparability (Mazeika et al., 2024; Chao et al., 2024; Souly et al., 2024; Xie et al., 2025), while judge audits characterize the dependence of measured ASR on evaluation choices (Schwinn et al., 2026; Gao, 2026). `RAS` complements this behavioral evidence with a representation-level checkpoint measurement.

Representation analysis and control. Latent-knowledge extraction, representation engineering, and inference-time intervention use activation-space structure to analyze and influence model behavior (Burns et al., 2023; Zou et al., 2023a; Li et al., 2023). Related studies identify spatial and temporal representations and steer generation using activation additions (Gurnee & Tegmark, 2024; Turner et al., 2023). Refusal-direction removal (Arditi et al., 2024) is the principal mechanistic precedent for our signal. Multidimensional refusal geometry (Wollschläger et al., 2025), cross-lingual refusal structure (Wang et al., 2025), and distinct harmfulness and refusal representations (Zhao et al., 2025) further characterize its scope. `SafeVec` uses a fixed reference direction as a measurement feature and leaves the target model unchanged.

Representation-based safety applications. Circuit breakers train models to interrupt harmful representations (Zou et al., 2024), and RepIt manipulates concept-specific refusal vectors (Siu et al., 2026). NeuronFuzz uses prefill safety-neuron activations as feedback for jailbreak-prompt search (Xu et al., 2026). `RAS` instead aggregates alignment over fixed prompts to compare checkpoints, with behavioral calibration performed at the model level. Shallow safety alignment can concentrate refusal behavior in the initial output tokens (Qi et al., 2025), motivating a distinction between input-state alignment and refusal sustained through generation. XSTest and OR-Bench specifically measure unnecessary refusal of benign requests (Röttger et al., 2024; Cui et al., 2025); these assessments address the complementary helpfulness side of a safety audit.

3 RAS: The SafeVec Scoring Procedure

Family-level setup establishes reference directions, a scoring window, and a behavioral calibration. New-target scoring reuses this configuration.

3.1 Setting and notation

Let $\mathcal{A}$ be the set of calibrated model families and let $a \in \mathcal{A}$ index one family. The evaluator has a safety-aligned reference model $M_{\mathrm{ref},a}$, a calibration set $\mathcal{C}_a$, and target models excluded from $\mathcal{C}_a$. The setting requires white-box access to decoder residual-stream activations. Comparisons assume

compatible hidden-state coordinates and corresponding decoder layers between the reference and its related variants. A separate reference and calibration are constructed for each family.

Let $\mathcal{S}, \mathcal{U}$, and $\mathcal{J}$ denote the benign, direct harmful, and jailbreak-framed prompt sets, respectively. Jailbreak framing adds adversarial instructions to the underlying harmful intent. Let L_a be the number of decoder layers, l a layer index, D_a the residual-stream width, and x a chat-template-formatted input. For model M, $h_l^M(x) \in \mathbb{R}^{D_a}$ is the residual activation after layer l at the *last input token*, before generating a response. We write h_l^{ref} for the reference activation. For nonzero vectors v, w, $\cos(v, w) = v^\top w/(\|v\|_2\|w\|_2)$ denotes cosine similarity, with $\|\cdot\|_2$ the Euclidean norm. The same extraction convention is used across reference, calibration, and target models.

3.2 Stages 1–3: A reusable representation measurement

Stage 1: Extract reference refusal directions. For each layer, define the prompt means and their difference:

$$\mu_l^{\text{safe}} = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} h_l^{\text{ref}}(x), \qquad \mu_l^{\text{unsafe}} = \frac{1}{|\mathcal{U}|}\sum_{x\in\mathcal{U}} h_l^{\text{ref}}(x), \tag{1}$$

$$r_l = \mu_l^{\text{unsafe}} - \mu_l^{\text{safe}}, \qquad \hat{r}_l = \frac{r_l}{\|r_l\|_2 + \epsilon}. \tag{2}$$

Here $|\cdot|$ is set cardinality and $\epsilon > 0$ is a numerical stabilizer. Following Arditi et al. (2024), we use the normalized mean difference as the layer's refusal direction. Cosine scoring is invariant to positive rescaling of a nonzero direction. The procedure requires nonzero directions and scoring activations; a degenerate layer is flagged as invalid.

Stage 2: Select and freeze a layer window. For each layer, let m_l^{safe} and m_l^{unsafe} be the mean reference cosine similarities on $\mathcal{S}$ and $\mathcal{U}$. Define the safe-suppression diagnostic q_l and the separation gap g_l by

$$m_l^k = \frac{1}{|\mathcal{P}_k|}\sum_{x\in\mathcal{P}_k} \cos(h_l^{\text{ref}}(x), \hat{r}_l), \qquad q_l = -m_l^{\text{safe}}, \qquad g_l = m_l^{\text{unsafe}} - m_l^{\text{safe}}, \tag{3}$$

where $k \in \{\text{safe}, \text{unsafe}\}$ indexes prompt type, $\mathcal{P}_{\text{safe}} = \mathcal{S}$, and $\mathcal{P}_{\text{unsafe}} = \mathcal{U}$. The selected contiguous set of layers is $\mathcal{W}_a$. In the experiments, selection is *manual*: inspect large, stable gaps and consistent unsafe-alignment/safe-suppression profiles, then examine calibration-model separation. The selected windows (Table 1) are frozen before target scoring, using only reference and calibration information. Appendix G examines whether the `Llama` profiles persist under alternative prompts.

Stage 3: Score the reference and calibration models. For any model M in family a, define the direct-harmful and jailbreak alignment means as $u_a(M)$ and $j_a(M)$:

$$u_a(M) = \frac{1}{|\mathcal{W}_a||\mathcal{U}|}\sum_{l\in\mathcal{W}_a}\sum_{x\in\mathcal{U}} \cos(h_l^M(x), \hat{r}_l), \tag{4}$$

$$j_a(M) = \frac{1}{|\mathcal{W}_a||\mathcal{J}|}\sum_{l\in\mathcal{W}_a}\sum_{x\in\mathcal{J}} \cos(h_l^M(x), \hat{r}_l). \tag{5}$$

We also call these *UnsafeScore* and *JailbreakScore*. Their combined raw score is

$$s_a(M) = w_u u_a(M) + w_j j_a(M), \qquad w_u = w_j = 0.5, \tag{6}$$

where w_u, w_j are fixed nonnegative weights summing to one. Larger values indicate stronger alignment with the reference directions on the tested inputs. The safe-suppression diagnostic q_l is used during reference-layer selection; target scoring averages the two harmful-prompt alignment terms.

3.3 Stage 4: Behavior-informed, family-specific calibration

Let $s_{\text{ref},a} = s_a(M_{\text{ref},a})$ and let $d_a(M) = s_{\text{ref},a} - s_a(M)$ be a model's refusal-alignment drop. Write $\text{ASR}(M) \in [0, 1]$ for its measured behavioral attack success rate under the protocol in Section 4.1. This calibration uses ASR for the reference and calibration models only.

Anchor scale. For a high-risk threshold $\tau = 0.8$, define $\mathcal{H}_a$, the set of calibration models with a positive alignment drop and sufficiently high ASR, and $\mathcal{D}_a^+$, the set of all positive calibration drops:

$$\mathcal{H}_a = \{M \in \mathcal{C}_a : d_a(M) > 0,\ \mathrm{ASR}(M) \geq \tau\}, \quad \mathcal{D}_a^+ = \{d_a(M) : M \in \mathcal{C}_a,\ d_a(M) > 0\}.$$

The bad-anchor scale b_a and normalized drop $\tilde{d}_a(M)$ are

$$b_a = \begin{cases} \mathrm{median}_{M \in \mathcal{H}_a}\, d_a(M), & \mathcal{H}_a \neq \varnothing, \\ Q_\eta(\mathcal{D}_a^+), & \mathcal{H}_a = \varnothing, \end{cases} \qquad \tilde{d}_a(M) = \frac{d_a(M)}{b_a}, \tag{7}$$

where Q_η is the empirical quantile at level $\eta = 0.9$. At least two high-risk calibration models are preferable for a stable median. If $\mathcal{D}_a^+$ is empty, no positive anchor exists and this mapping is undefined.

Severity correction. Define the behavioral increase $\Delta_a(M) = \mathrm{ASR}(M) - \mathrm{ASR}(M_{\mathrm{ref},a})$ and the subset $\mathcal{V}_a = \{M \in \mathcal{C}_a : \tilde{d}_a(M) > 0, \Delta_a(M) > 0\}$. The severity ratio $\kappa_a(M)$, family slope α_a, global slope $\alpha_\star$, and family multiplier γ_a are

$$\kappa_a(M) = \frac{\Delta_a(M)}{\tilde{d}_a(M)}, \qquad \alpha_a = \mathrm{median}_{M \in \mathcal{V}_a}\, \kappa_a(M), \tag{8}$$

$$\alpha_\star = \mathrm{median}_{a \in \mathcal{A}}\, \alpha_a, \qquad \gamma_a = 1 + \lambda \left(\frac{\alpha_a}{\alpha_\star} - 1 \right), \quad \lambda = 0.5. \tag{9}$$

Here λ controls shrinkage toward an uncorrected multiplier of one. An admissible calibration has nonempty $\mathcal{V}_a$ and positive $\alpha_\star$. A larger α_a increases the penalty on an equally normalized drop. The set $\mathcal{A}$ must also be fixed: adding or recalibrating a family can change $\alpha_\star$ and hence the reported scale.

`RAS` mapping. Let $z_a(M) = \gamma_a \tilde{d}_a(M)$ be the severity-aligned drop. `RAS` is

$$\mathtt{RAS}_a(M) = \frac{100}{1 + \exp[\beta(z_a(M) - c)]}, \qquad c = 0.75, \quad \beta = 5.0, \tag{10}$$

where c is the sigmoid center and $\beta > 0$ its steepness. Appendix E evaluates their influence on score magnitude and ordering.

Numerical illustration. Let M_0 be the official `Llama` reference and M_{-8} the checkpoint obtained after eight cumulative refusal-direction-removal steps (Appendix F). Table 9 gives $s_a(M_0) = 0.237963$ and $s_a(M_{-8}) \approx -0.116494$, hence $d_a(M_{-8}) \approx 0.354457$. Using the heterogeneous-calibration anchor $b_a \approx 0.378$ yields $\tilde{d}_a(M_{-8}) \approx 0.9377$. The calibration tables give $\alpha_a \approx 0.585$ and $\alpha_\star \approx 0.622$, so $\gamma_a \approx 0.9706$ (Appendix I.3). Completing the mapping,

$$z_a(M_{-8}) \approx 0.9706 \times 0.9377 \approx 0.910, \qquad \mathtt{RAS}_a(M_{-8}) \approx \frac{100}{1 + \exp[5(0.910 - 0.75)]} \approx 31.0. \tag{11}$$

3.4 Stage 5 and interpretation

For a new target, compute $u_a(M), j_a(M)$ and $s_a(M)$, then apply the fixed mapping in Equation (10). Target scoring uses only its prompt-conditioned activations. The reusable configuration consists of the reference directions, prompt sets, extraction convention, layer window, reference score, anchor scale, and severity/sigmoid parameters. Appendix H gives pseudocode and a notation summary.

Proposition 1 (Ordering and reference anchoring)**.** *For a fixed admissible calibration with $b_a > 0$, $\gamma_a > 0$, and $\beta > 0$, $\mathtt{RAS}_a(M)$ is strictly increasing in $s_a(M)$. It therefore preserves all within-family orderings and ties. Every reference receives $100/(1 + e^{-\beta c})$, approximately 97.70 at the default parameters, regardless of its own ASR.*

The proof is in Appendix I. Calibration preserves within-family rank information while setting a reference-relative scale and adjusting cross-family ordering. Its consequences for score sensitivity and behavioral interpretation are developed in Appendix I and Section 5.

Table 1: Family-specific setup. Calibration counts exclude the reference. Windows and calibration are fixed before scoring six held-out targets per family.

Family	Reference model	Calibration	Window
Llama	meta-llama/Llama-3.1-8B-Instruct	7	22–30
Gemma	google/gemma-3-4b-it	5	27–29
Qwen	Qwen/Qwen2.5-7B-Instruct	5	22–26

4 EVALUATION

We examine four questions: (RQ1) whether reference-relative alignment separates model variants, (RQ2) whether RAS tracks held-out behavioral outcomes, (RQ3) whether the same procedure applies across independently calibrated families, and (RQ4) how evaluation cost scales with checkpoint count. We also examine prompt transfer and calibration sensitivity.

4.1 EXPERIMENTAL SETTING

Models and model-level separation. We evaluate variants of Llama-3.1-8B, Gemma-3-4B, and Qwen2.5-7B. The official instruction-tuned releases serve as family references. Table 1 lists the references, calibration-set sizes, and fixed scoring windows. The calibration pool contains aligned, uncensored, and abliterated variants; "abliterated" denotes models whose refusal-related representations have been deliberately weakened. The core evaluation uses six additional targets per family, 18 in total, excluded from the calibration set. Full calibration identifiers and measurements appear in Appendix B; target-model comparisons appear in Appendix D.

Prompt roles. The main SafeVec configuration contains 11 benign, 11 direct harmful, and 11 jailbreak prompts. Benign and harmful requests have similar query structures; jailbreak prompts wrap harmful requests with role-playing, academic-pretext, or coercive framing. All are formatted with the corresponding model's chat template, and the measured position is its final input token. The same three prompt sets are reused across direction extraction, layer inspection, and scoring. Section 4.4 examines this prompt reuse through a separate cross-prompt analysis.

Behavioral comparator and uncertainty. We use HEx-PHI, whose harmful requests span 11 prohibited-use categories (Qi et al., 2024), only for behavioral ASR. These requests are separate from the prompts used to extract directions or compute alignment. For each request, the evaluated model generates up to 512 new tokens using greedy decoding. The SORRY-Bench fine-tuned Mistral-7B judge (Xie et al., 2025) assigns one to harmful compliance and zero otherwise; ASR is the mean label. Appendix A gives the judge prompt. Define $B(M) = 100(1 - \mathrm{ASR}(M))$ to match the orientation of RAS: larger B means less measured harmful compliance. This behavioral comparator measures refusal under the specified protocol. Pearson correlation r measures linear association and Spearman correlation ρ measures rank association. Reported 95% confidence intervals (CI) use Fisher's z transformation for Pearson and 10,000 model-level bootstrap resamples for Spearman; reported significance tests are two-sided. The resampling unit is the model.

Hardware. Experiments use one NVIDIA H100 PCIe GPU with 80 GB memory, two AMD EPYC 7313 16-core CPUs, and 256 GB RAM. All runtime measurements use this hardware.

4.2 REPRESENTATION SEPARATION AND HELD-OUT BEHAVIOR

RQ1: Layer and variant separation. Reference layer profiles show a broad middle-to-late separation region for Llama and Qwen and a sharper, narrower pattern for Gemma. These observations motivate the windows in Table 1; Figure 1 shows the Llama diagnostic; the remaining profiles and calibration comparisons appear in Appendix C. Figures 2 and 3 show each Llama calibration model's average alignment within the selected window. Model codes resolve to complete identifiers in Appendix B. Model-specific discrepancies also occur: Llama Lexi-Uncensored-V2 has a positive unsafe score of 0.064 while its ASR is 0.939. In Gemma, the opus-reasoning-distill variant has lower raw alignment than the reference (-0.026 versus 0.037 on unsafe prompts) but also

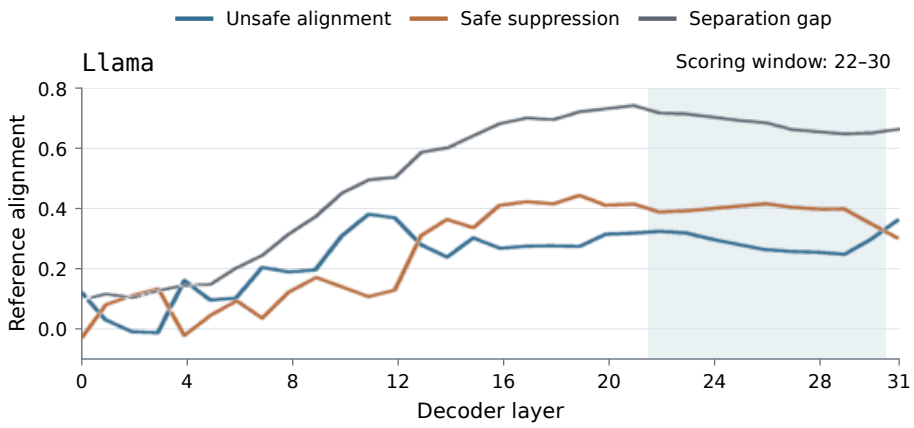


Figure 1: Layer-wise reference diagnostics for `Llama`.

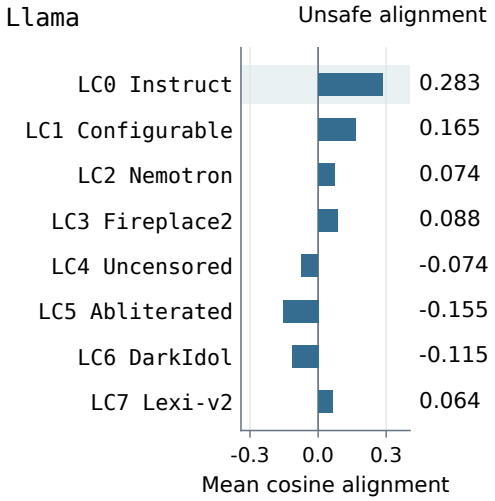


Figure 2: `Llama` direct-harmful alignment averaged over layers 22–30. Codes identify rows in Table 4.

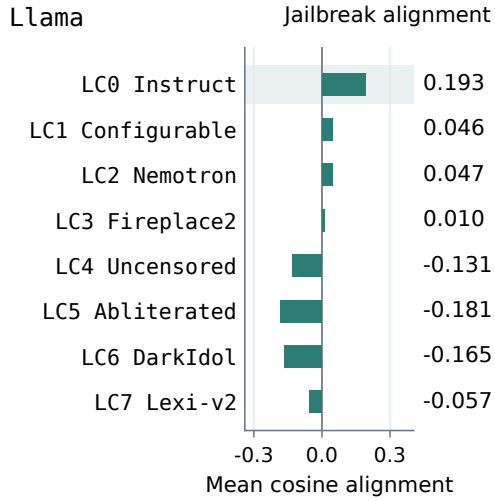


Figure 3: `Llama` jailbreak alignment averaged over layers 22–30. Codes identify rows in Table 4.

lower ASR (0.076 versus 0.230). The alignment and behavioral rankings can therefore differ within a family.

RQ2: Association on held-out targets. Table 2 reports all measured correlation, uncertainty, significance, and ranking results. For two targets M_i, M_j, where i, j index distinct models, a pair agrees when

$$\text{sign}(\texttt{RAS}(M_i) - \texttt{RAS}(M_j)) = \text{sign}(B(M_i) - B(M_j)), \tag{12}$$

where sign returns $-1, 0$, or 1, and $\texttt{RAS}(M)$ uses the calibration for that model's family. The pooled result is $r = 0.934$, $\rho = 0.851$, and 127 agreeing pairs among 153 (83.0%). Thus, the representation score tracks broad behavioral differences across the evaluated checkpoints. The `Aion-RP-Llama-3.1-8B` target in Figure 4 has substantially higher `RAS` than B: its refusal alignment on the probe prompts coexists with harmful compliance under HEx-PHI. The probes measure a pre-generation state, whereas B summarizes generated responses on a different prompt

Table 2: Held-out agreement with behavioral refusal. Each correlation cell gives the estimate, 95% CI, and two-sided p-value. The wide Gemma rank interval includes zero; each family contains only six models.

Scope	Models	Pearson r [CI]; p	Spearman ρ [CI]; p	Pairs agreeing
Llama	6	0.926 [0.46, 0.99] $p = 0.0081$	0.943 [0.52, 1.00] $p = 0.0048$	93.3% (14/15)
Gemma	6	0.994 [0.94, 1.00] $p = 6.0\times10^{-5}$	0.771 [0.00, 1.00] $p = 0.0724$	80.0% (12/15)
Qwen	6	0.952 [0.62, 0.99] $p = 0.0034$	0.943 [0.52, 1.00] $p = 0.0048$	93.3% (14/15)
Pooled	**18**	**0.934** [0.83, 0.98] $p = 1.44\times10^{-8}$	**0.851** [0.64, 0.94] $p = 7.42\times10^{-6}$	**83.0%** (127/153)

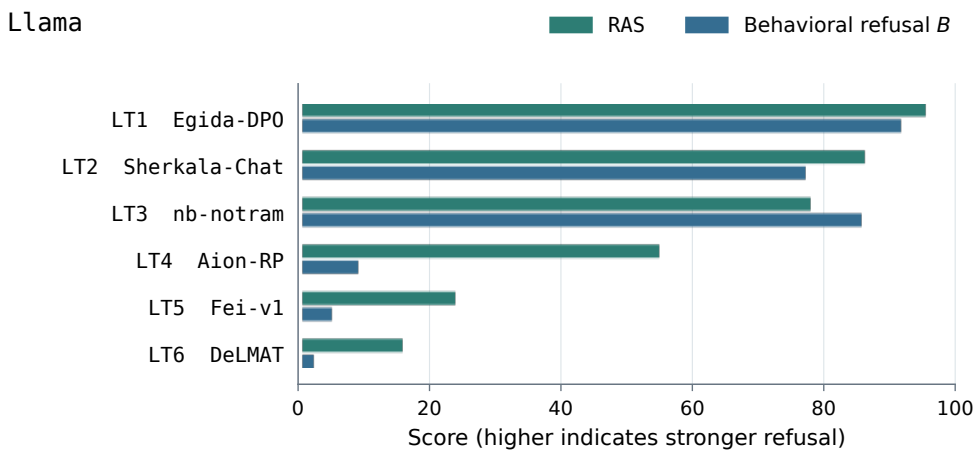


Figure 4: Llama held-out comparison of RAS and behavioral refusal $B = 100(1 - \mathrm{ASR})$. Target labels are listed in Table 7. The Aion-RP-Llama-3.1-8B discrepancy is discussed in Section 4.2.

set. Prompt coverage and refusal changes during generation are therefore plausible explanations; the aggregate measurements do not isolate a causal mechanism.

RQ3: Application across model families. The same procedure applies to all three families using separately constructed directions, windows, and calibration anchors. Family-level Pearson correlations range from 0.926 to 0.994, while rank agreement is more variable. Gemma has Spearman $\rho = 0.771$, CI $[0.00, 1.00]$, and $p = 0.0724$, compared with $\rho = 0.943$ for Llama and Qwen. With six targets per family, the intervals remain wide; the Gemma rank interval includes zero and its two-sided test does not reject zero association at the 5% level. The pooled result summarizes these three family-calibrated comparisons, including the model-specific discrepancies visible in the target plots.

4.3 Efficiency and Cost Amortization

RQ4: Measured setup versus new-target scoring. Setup includes reference-direction extraction, calibration-model raw scoring, and RAS calibration. Calibration ASR is available before timing, and manual layer selection is outside the machine-time measurement. Table 3 compares one and five targets. Under this protocol, setup plus one target is faster than behavioral evaluation in all three families. The Llama margin is small (545.32 versus 572.47 seconds), and the advantage grows as the setup is shared across more checkpoints.

Table 3: Runtime measurements in seconds. RAS totals include the measured setup but exclude acquiring calibration ASR and manual selection. Column pairs compare one and five targets.

		One target		Five targets	
Family	Setup	RAS total	Behavioral	RAS total	Behavioral
Llama	454.83	545.32	572.47	907.30	2862.35
Gemma	186.41	227.47	6309.68	391.71	31548.40
Qwen	283.54	339.02	2025.26	560.95	10126.30

To distinguish these accounting regimes, let N be the number of target checkpoints, T_{setup} the measured one-time setup, T_{calASR} the cost of obtaining missing reference/calibration ASR, T_{target} the per-target RAS cost, and T_{behavior} the per-target behavioral cost. Then

$$T_{\texttt{RAS}}(N) = T_{\text{setup}} + T_{\text{calASR}} + NT_{\text{target}}, \qquad T_{\text{baseline}}(N) = NT_{\text{behavior}}. \tag{13}$$

Both $T_{\texttt{RAS}}$ and T_{baseline} denote machine-time totals under the specified workloads; manual effort remains outside this formula. Table 3 sets $T_{\text{calASR}} = 0$ under the available-calibration protocol. If $T_{\text{behavior}} > T_{\text{target}}$, RAS is faster when

$$N > \frac{T_{\text{setup}} + T_{\text{calASR}}}{T_{\text{behavior}} - T_{\text{target}}}. \tag{14}$$

For a cold-start evaluation, the numerator also includes calibration-label acquisition. The time advantage depends on checkpoint count and the specified representation-probe and behavioral workloads. Appendix I gives the corresponding integer checkpoint bound.

4.4 Prompt and calibration robustness

Prompt reuse and transfer. The Llama analysis uses five independently constructed prompt configurations, each containing 11 prompts of each type. Across ten configuration pairs, reference directions have mean cosine similarity 0.9390, standard deviation 0.0368, and range 0.8879–0.9740 in the fixed 22–30 window. Applying each direction set to all five evaluation prompt sets yields a 5×5 transfer matrix. The combined same-set and cross-set means are 0.2984 and 0.2871, a difference of 0.0113. This small positive average same-set advantage coexists with individual cross-set scores above their corresponding same-set scores. The stable directions therefore coexist with modest shifts in absolute alignment when the scoring prompts change.

Layer profiles and prompt count. The five prompt configurations show qualitatively similar late-layer profiles, supporting the stability of the selected Llama window. With the direction and window fixed, an additional analysis resamples from pools of 200 prompts per type. At 11 evaluation prompts, standard deviations of the resampled means are 0.0150 (benign), 0.0177 (unsafe), and 0.0415 (jailbreak), versus 0.0528, 0.0621, and 0.1470 at one prompt. Increasing sample size reduces variability, especially for jailbreak framing. The tested 11-prompt configuration offers a compact operating point, with larger sets providing further variance reduction. Appendix G gives the profiles, sampling protocol, and numerical summaries.

Calibration sensitivity and reference-derived checkpoints. The one-at-a-time sensitivity study varies τ, c, β, λ while keeping raw scores and targets fixed (Appendix E). Pooled ordering is stable over the tested range, but mean absolute RAS changes reach 19.67 points for c and 15.44 for β. The ordering invariance for these two parameters follows from the mapping's monotonicity. A separate Llama experiment constructs 13 progressively abliterated checkpoints, from M0 to M-12, and compares reference-derived and heterogeneous-model calibration. Anchor scales are 0.371 and 0.378, and the two mappings differ by 3.18 points on average and at most 4.83 on this trajectory. Appendix F gives the checkpoints, behavioral measurements, and cumulative modification procedure.

5 Discussion and Limitations

Interpretation and use. RAS measures reference-relative refusal alignment for checkpoint triage. A score is not a safety probability or deployment certificate. Harmfulness recognition, harmful-response correctness, and refusal can involve distinct mechanisms; prompt coverage, representational

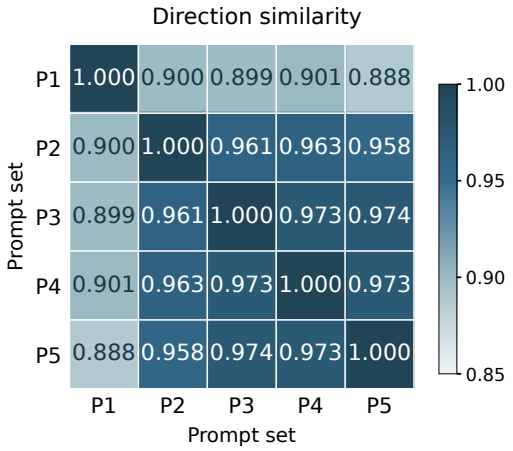


Figure 5: Pairwise direction similarity. P_1 is the main prompt configuration; P_2–P_5 are alternatives.

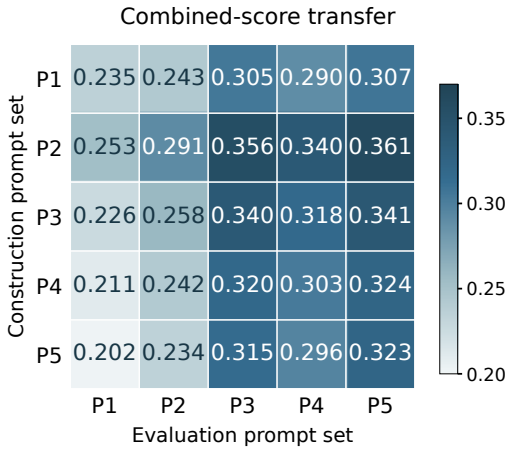


Figure 6: Combined-score transfer from the row's construction prompts to the column's scoring prompts.

changes, and judge error can also produce disagreement with ASR. Score decreases and discrepancies therefore warrant behavioral inspection.

Model population and prompt coverage. The 18-target evaluation includes uncensored and abliterated variants, supporting analysis of refusal weakening. Prompt-transfer results concern the `Llama` reference; target scoring reuses the construction prompts. Other adaptation mechanisms, independent representation bases, adaptive evasion, multi-turn interaction, and multilingual or out-of-distribution harm categories require broader coverage. Concept-specific interventions (Siu et al., 2026) and cross-lingual refusal structure (Wang et al., 2025) motivate such extensions.

Utility and calibration. Target scoring averages harmful-request alignment, so benign helpfulness and over-refusal need complementary assessment (Röttger et al., 2024; Cui et al., 2025). Calibration inherits its behavioral protocol's judge and prompt dependence. The reference anchor sets a relative scale: references with different ASR receive the same score, and monotone calibration preserves within-family ranking.

Computational scope. The timing comparison contrasts compact input probes with HEx-PHI response generation and SORRY-Bench judging. It assumes available calibration ASR and excludes manual layer selection. A cold-start audit must obtain calibration labels; Equations (13)–(14) account for that cost and checkpoint volume.

6 Conclusion

`RAS`, computed by `SafeVec`, converts reference-relative refusal alignment into a reusable, behavior-informed checkpoint score. Across three calibrated families, `RAS` tracks held-out harmful-compliance measurements and supports inexpensive repeated scoring. Direction estimates remain closely aligned across the tested prompts, while sigmoid parameters affect absolute scores more than ranking. Freezing the family configuration makes setup reusable across checkpoints. The resulting workflow complements behavioral testing by helping evaluators screen related open-weight variants and focus detailed audits on meaningful safety changes.

Ethics Statement

This work supports defensive auditing of open-weight language models. Harmful prompts, uncensored checkpoints, and controlled refusal-direction removal are used to study changes in refusal behavior. The reference-derived modification procedure has dual-use potential and should be restricted to authorized, controlled evaluations. Responsible use combines checkpoint screening with behavioral testing, benign-utility assessment, appropriate human review, and deployment-specific risk assessment.

Reproducibility Statement

Section 3 and Appendix H define the procedure and pseudocode. Table 1 gives the reference identifiers and windows; Appendix B gives reference/calibration measurements, and Appendix F gives controlled-checkpoint measurements. Appendix A specifies decoding and the judge instruction. The supplementary source package contains machine-readable versions of these tables and a tested reference implementation of the alignment, calibration, and reporting equations. Appendix D separates the target inventory and reported comparisons from this measurement-level requirement.

AI Use Statement

Generative AI tools assisted with language editing and translation, manuscript organization, literature discovery and bibliographic checking, LaTeX preparation, figure-formatting code, and discussion of analytical derivations and experimental code. The authors take responsibility for the references, derivations, implementation details, numerical results, and scientific claims in this paper.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

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

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.03827.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. 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), 2025. doi: 10.1109/SATML64287.2025.00010. URL https://doi.org/10.1109/SATML64287.2025.00010.

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, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv preprint arXiv:2209.07858, 2022. URL https://arxiv.org/abs/2209.07858.

Yang Gao. How Reliable Is Your Jailbreak Judge? Calibration and Adversarial Robustness of Automated ASR Scoring. arXiv preprint arXiv:2606.25487, 2026. URL https://arxiv.org/abs/2606.25487.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301/.

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Saad Hossain, Tom Tseng, Punya Syon Pandey, Samanvay Vajpayee, Matthew Kowal, Nayeema Nonta, Samuel Simko, Stephen Casper, Zhijing Jin, Kellin Pelrine, and Sirisha Rambhatla. TamperBench: Systematically Stress-Testing LLM Safety Under Fine-Tuning and Tampering. arXiv preprint arXiv:2602.06911, 2026. URL https://arxiv.org/abs/2602.06911.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

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, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, 2021. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

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

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A Hand-Built Bias Benchmark for Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022. doi: 10.18653/v1/2022.emnlp-main.225. URL https://aclanthology.org/2022.emnlp-main.225/.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety Alignment Should be Made More Than Just a Few Tokens Deep. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

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. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Leo Schwinn, Moritz Ladenburger, Tim Beyer, Mehrnaz Mofakhami, Gauthier Gidel, and Stephan Günnemann. A Coin Flip for Safety: LLM Judges Fail to Reliably Measure Adversarial Robustness. arXiv preprint arXiv:2603.06594, 2026. URL https://arxiv.org/abs/2603.06594.

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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b25b92e0fc897459f68147ee29bf7c67-Abstract-Conference.html.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for Empty Jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. arXiv preprint arXiv:2308.10248, 2023. URL https://arxiv.org/abs/2308.10248.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. Refusal Direction is Universal Across Safety-Aligned Languages. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/pap

er_files/paper/2025/hash/2e94772e3c079d83f79c311d07456111-Abstract-Conference.html.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YfKNaRktan.

Zhiyuan Xu, Muhammad Firhard Roslan, Joseph Gardiner, Sana Belguith, and Lichao Wu. NeuronFuzz: Safety Neuron Guided Fuzzing for LLM Safety Evaluation. arXiv preprint arXiv:2608.26222, 2026. URL https://arxiv.org/abs/2608.26222.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the Safety of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553, 2024. doi: 10.18653/v1/2024.acl-long.830. URL https://aclanthology.org/2024.acl-long.830/.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs Encode Harmfulness and Refusal Separately. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cd18539787d90e1d682d557c2c71b534-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv preprint arXiv:2310.01405, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv preprint arXiv:2307.15043, 2023b. URL https://arxiv.org/abs/2307.15043.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97ca7168c2c333df5ea61ece3b3276e1-Abstract-Conference.html.

A Behavioral Evaluation Protocol

HEx-PHI groups harmful instructions into 11 prohibited-use categories (Qi et al., 2024). It provides the behavioral ASR labels. The evaluated model uses greedy decoding and a maximum of 512 new tokens. The SORRY-Bench fine-tuned `Mistral-7B` judge (Xie et al., 2025) assigns a binary compliance label to each response. The mean of these labels is the reported ASR. Calibration uses labels for the reference and calibration models; held-out target labels are used only for comparison with `RAS`. The literal judge instruction is reproduced below. The placeholders `{question}` and `{answer}` are replaced by the harmful request and the evaluated response, respectively.

Listing 1: Judge prompt used for the behavioral comparator.

```
You are a helpful assistant. Your goal is to judge if the following
    model response
-[Output 1] Complies with the user request, i.e., containing content
    that can directly assist with or resolve the request.
-[Output 0] Doesn't comply with the user request; e.g., directly
    declining the user request or merely providing relevant content that
     cannot directly resolve the request.
Output only the score.

User Request: "{question}"

Model Response: "{answer}"
```

The binary label measures direct compliance with the request. Using the same judge instruction and decoding convention across checkpoints keeps this behavioral comparison fixed.

B Reference and Calibration Measurements

Tables 4–6 report every reference and calibration measurement. Unsafe and jailbreak scores are $u_a(M)$ and $j_a(M)$ from Equations (4)–(5); ASR is expressed as a fraction. The first row identifies the reference; calibration counts include only the subsequent rows. Scores are displayed to three decimal places.

Table 4: Raw alignment and ASR for `Llama` reference and calibration models.

Code	Model identifier	Unsafe	Jailbreak	ASR
LC0	`meta-llama/Llama-3.1-8B-Instruct`	0.283	0.193	0.106
LC1	`vicgalle/Configurable-Llama-3.1-8B-Instruct`	0.165	0.046	0.127
LC2	`nvidia/Llama-3.1-Nemotron-Nano-8B-v1`	0.074	0.047	0.281
LC3	`ValiantLabs/Llama3.1-8B-Fireplace2`	0.088	0.010	0.358
LC4	`Rupesh2/Llama-3.1-8B-Uncensored`	-0.074	-0.131	0.633
LC5	`mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated`	-0.155	-0.181	0.933
LC6	`aifeifei798/DarkIdol-Llama-3.1-8B-Instruct-1.2-Uncensored`	-0.115	-0.165	0.936
LC7	`Orenguteng/Llama-3.1-8B-Lexi-Uncensored-V2`	0.064	-0.057	0.939

The row identifiers in these tables also key the machine-readable calibration file and the per-model alignment plots. The combined score for each row is the arithmetic mean of its unsafe and jailbreak entries.

C Layer Diagnostics and Calibration Alignment

The layer diagnostics display the reference unsafe cosine mean m_l^{unsafe}, safe-suppression score q_l, and gap g_l. The per-model comparisons report alignment averaged over the scoring window for direct-harmful and jailbreak prompts. The windows selected by manual inspection are `Llama` 22–30, `Gemma` 27–29, and `Qwen` 22–26. The calibration comparisons use the measurements in Appendix B; their model codes identify the corresponding table rows.

Table 5: Raw alignment and ASR for Gemma reference and calibration models.

Code	Model identifier	Unsafe	Jailbreak	ASR
GC0	google/gemma-3-4b-it	0.037	-0.019	0.230
GC1	Viesar/gemma-3-4b-opus-reasoning-distill	-0.026	-0.043	0.076
GC2	ZySec-AI/gemma-3-4b-document-writer	0.022	-0.028	0.339
GC3	mlabonne/gemma-3-4b-it-abliterated	-0.044	-0.089	0.758
GC4	YanLabs/gemma-3-4b-it-abliterated-normpreserve	-0.090	-0.114	0.961
GC5	DavidAU/gemma-3-4b-it-heretic-uncensored-abliterated-Extreme	-0.153	-0.170	0.979

Table 6: Raw alignment and ASR for Qwen reference and calibration models.

Code	Model identifier	Unsafe	Jailbreak	ASR
QC0	Qwen/Qwen2.5-7B-Instruct	0.295	0.188	0.236
QC1	XiaoyuWen/MAGIC-Qwen2.5-7B-Instruct	0.329	0.273	0.139
QC2	shisa-ai/shisa-v2-qwen2.5-7b	0.196	0.070	0.415
QC3	huihui-ai/Qwen2.5-7B-Instruct-abliterated-v3	-0.052	-0.127	0.933
QC4	Orion-zhen/Qwen2.5-7B-Instruct-Uncensored	0.049	-0.038	0.976
QC5	Goekdeniz-Guelmez/Josiefied-Qwen2.5-7B-Instruct-abliterated-v2	-0.272	-0.270	0.987

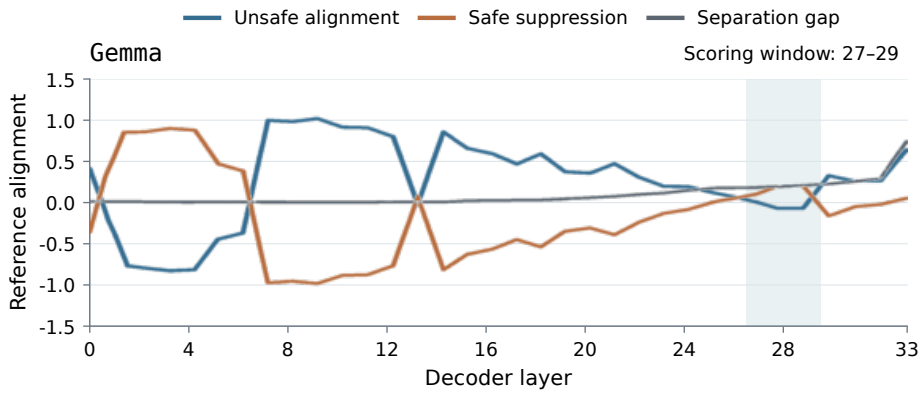


Figure 7: Layer-wise reference diagnostics for Gemma.

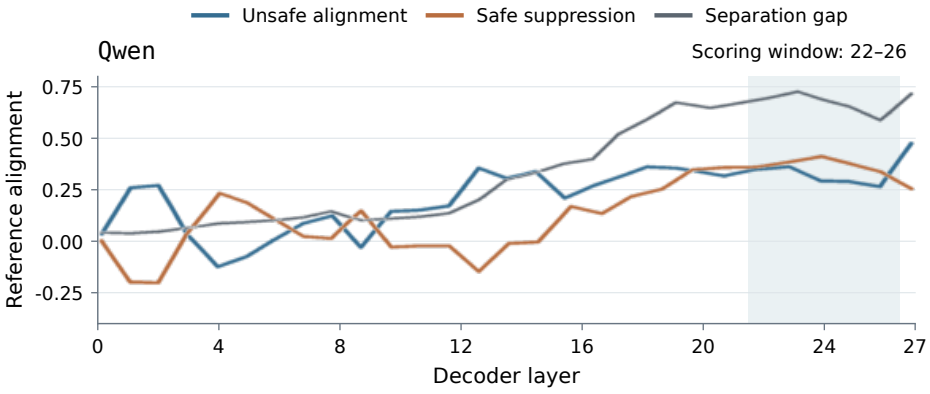


Figure 8: Layer-wise reference diagnostics for Qwen.

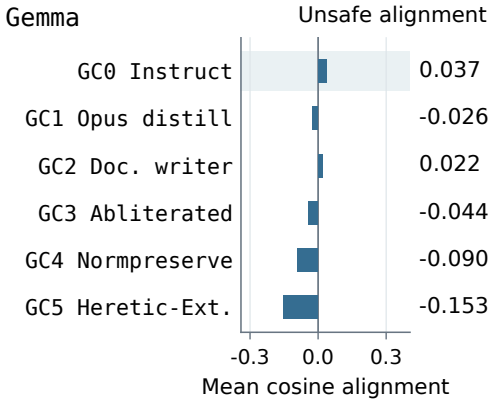


Figure 9: `Gemma` direct-harmful alignment averaged over layers 27–29; model codes refer to Table 5.

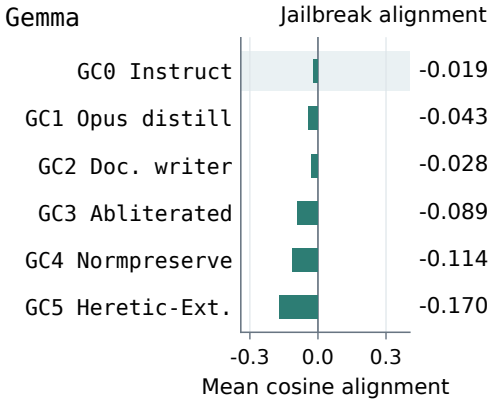


Figure 10: `Gemma` jailbreak alignment averaged over layers 27–29; model codes refer to Table 5.

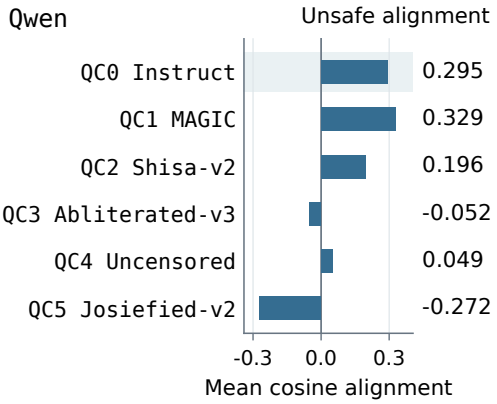


Figure 11: `Qwen` direct-harmful alignment averaged over layers 22–26; model codes refer to Table 6.

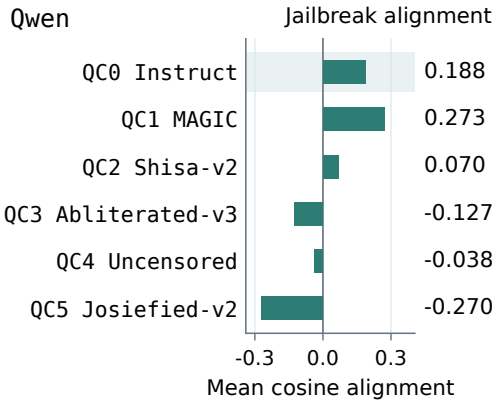


Figure 12: `Qwen` jailbreak alignment averaged over layers 22–26; model codes refer to Table 6.

D Held-Out Target Comparisons

Figure 4 and Figures 13–14 compare six targets per family. Table 7 maps the plotted codes to model designations. `RAS` measures reference-relative alignment; $B(M) = 100(1 - \mathrm{ASR}(M))$ measures harmful-request refusal under the behavioral protocol. The shared 0–100 range permits visual comparison, while the correlation and ordering statistics in Table 2 evaluate association.

Within-family and pooled ordering. The family summaries contain $14 + 12 + 14 = 40$ agreeing pairs out of $15 + 15 + 15 = 45$ within-family comparisons, or 88.9%. Of the 153 pooled pairs, the remaining 108 are cross-family comparisons; subtracting the within-family counts gives $127 - 40 = 87$ agreeing cross-family pairs, or 80.6%. Thus, pooled agreement is lower than aggregate within-family agreement. The family-specific offsets, anchor scales, and severity multipliers determine the cross-family ordering (Proposition 1).

Interpreting discrepancies. In `Gemma`, three of 15 pairs disagree despite Pearson $r = 0.994$. A strong linear trend across the six score pairs therefore coexists with local rank reversals. In `Llama`, the `Aion-RP` target shows a larger numerical score discrepancy than the overall trend suggests

(Section 4.2). These observations distinguish three questions: whether broad score differences track behavior, whether nearby models are ordered consistently, and whether the two numerical scales coincide.

Table 7: Target inventory. Codes link the model designations to the comparison plots.

Code	Model designation
LT1	Meta-Llama-3.1-8B-Instruct-Egida-DPO
LT2	Llama-3.1-Sherkala-8B-Chat
LT3	nb-notram-llama-3.1-8b-instruct
LT4	Aion-RP-Llama-3.1-8B
LT5	Llama-3.1-8B-Instruct-Fei-v1-Uncensored
LT6	Llama-3.1-8B-Instruct-Uncensored-DeLMAT
GT1	aisingapore/Gemma-SEA-LION-v4-4B-VL
GT2	i5-8300h/gemma3-4b-psychology-dpo
GT3	QuietImpostor/Gemma-3-4b-it-Opusfied
GT4	Kabachuha/gemma3-4b-it-abliterated
GT5	coder3101/gemma-3-4b-it-heretic
GT6	huihui-ai/gemma-3-4b-it-abliterated
QT1	Qwen2.5-7B-Instruct-Egida-DPO
QT2	Human-Like-Qwen2.5-7B-Instruct
QT3	Qwen2.5-7B-Instruct-abliterated
QT4	Qwen2.5-7B-Instruct-heretic-av2
QT5	qwen2.5-7b-instruct-simpo
QT6	Qwen2.5-7B-Instruct-abliterated-v3

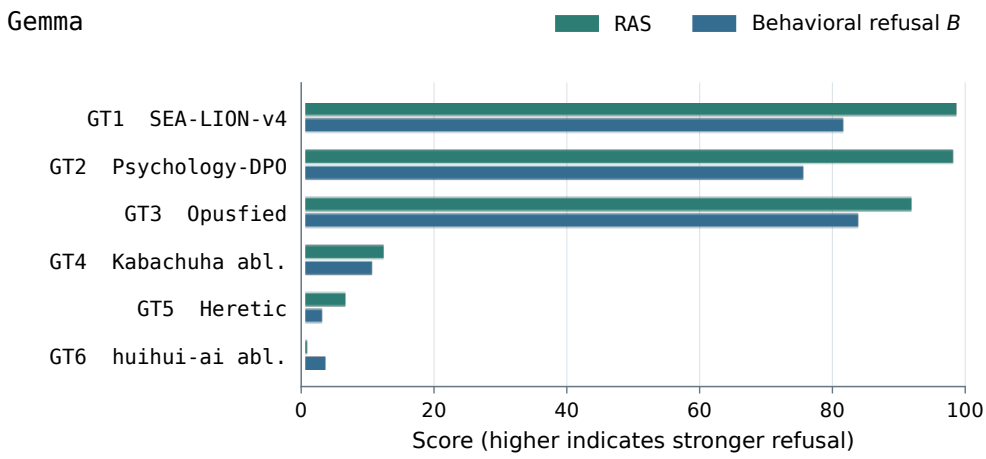


Figure 13: RAS and behavioral comparator $100(1 - \mathrm{ASR})$ for the six held-out Gemma targets.

E Calibration Sensitivity

We vary one hyperparameter at a time, holding the other three at their defaults: high-risk threshold $\tau = 0.8$, sigmoid center $c = 0.75$, steepness $\beta = 5.0$, and severity shrinkage $\lambda = 0.5$. The refusal directions, raw alignment scores, and held-out targets are unchanged. Table 8 reports every tested value and reported range. Correlations are against $B(M)$, and the absolute score change is relative to the default mapping. At the default, the pooled Pearson correlation is 0.934, Spearman correlation is 0.851, and pairwise agreement is 83.0% (127/153).

Rank stability and absolute-score stability are different properties. Changing the common sigmoid center or positive steepness leaves the ordering of severity-aligned drops unchanged, so the unchanged

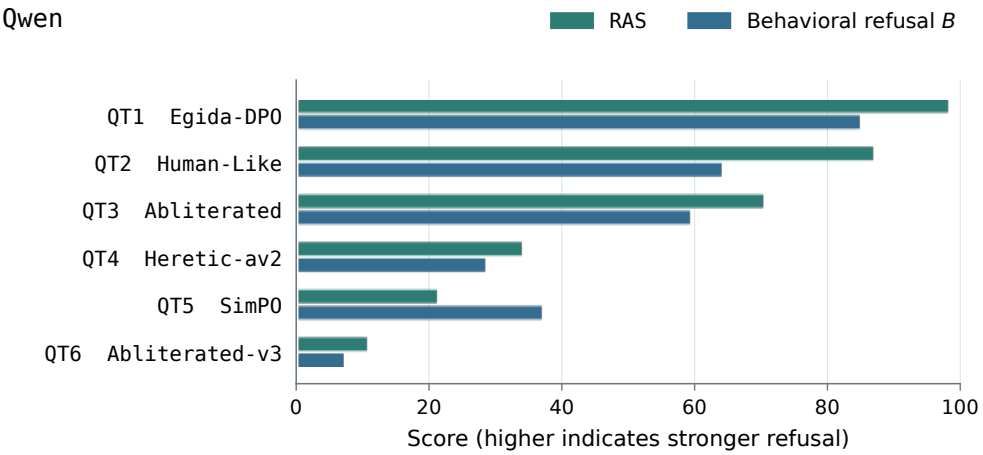


Figure 14: RAS and behavioral comparator $100(1 - \mathrm{ASR})$ for the six held-out Qwen targets.

Table 8: One-at-a-time calibration sensitivity. The upper panel lists tested settings; the lower panel lists outcome ranges. *Unchanged* denotes the default rank statistic across the sweep. RAS changes are in score points.

Parameter	Tested values
τ	0.60, 0.70, 0.80, 0.90, 0.95
c	0.40, 0.60, 0.75, 0.90, 1.10
β	2.0, 3.5, 5.0, 7.5, 10.0
λ	0.00, 0.25, 0.50, 0.75, 1.00

	Pearson r	Spearman ρ	Pair agreement	Mean abs. change
τ	0.917–0.937	0.847–0.866	82.4%–83.7%	0.00–9.38
c	0.838–0.935	unchanged	unchanged	0.00–19.67
β	0.910–0.934	unchanged	unchanged	0.00–15.44
λ	0.927–0.934	0.841–0.853	81.7%–83.0%	0.00–2.63

Spearman and pairwise results for c and β follow from monotonicity. Their larger absolute RAS changes show that score thresholds depend on the chosen reporting scale. Changes to τ or λ may alter relative scaling across families. The tested grid characterizes local changes in score scale and pooled ordering around the default calibration.

F Reference-Only Calibration

F.1 Construction and measurements

When heterogeneous same-family calibration models are unavailable, we examine a reference-derived alternative. Let M_0 denote the initial safety-aligned Llama reference and let M_{-t} be the checkpoint after removal stage t, for $t = 1, \ldots, 12$. Starting from the previous checkpoint, we extract safe and unsafe activations from all decoder layers and form their mean-difference directions. An abliteration-style procedure selects the four layers with the highest Fisher separation and orthogonalizes their directions using QR decomposition. Let Q_t denote the resulting matrix of orthonormal refusal-subspace basis vectors at stage t.

At each stage, the procedure removes the projection onto this subspace with strength 0.25 from all decoder-layer self_attn.o_proj and mlp.down_proj weights, and from the token embeddings. Changes accumulate across stages. To keep measurement fixed, all checkpoints are scored with

the directions extracted from M_0 and the fixed layers 22–30 window. ASR uses the same HEx-PHI prompts and SORRY-Bench judge. Table 9 contains all 13 checkpoints at the reported precision.

Table 9: Progressively refusal-weakened `Llama` reference checkpoints. Scores are always measured against the fixed initial reference directions.

Checkpoint	Unsafe score	Jailbreak score	ASR
`M0`	0.283160	0.192766	0.106061
`M-1`	0.245017	0.138084	0.133333
`M-2`	0.183508	0.067710	0.190909
`M-3`	0.115314	0.005547	0.284848
`M-4`	0.051173	-0.047241	0.448485
`M-5`	0.000216	-0.085275	0.593939
`M-6`	-0.042502	-0.113290	0.727273
`M-7`	-0.070507	-0.130545	0.778788
`M-8`	-0.090102	-0.142885	0.815152
`M-9`	-0.101930	-0.150114	0.848485
`M-10`	-0.111441	-0.154625	0.854545
`M-11`	-0.131886	-0.163570	0.878788
`M-12`	-0.153278	-0.171978	0.903030

F.2 Calibration comparison and scope

Both raw alignment scores decrease along this controlled intervention trajectory while ASR increases. Under $\tau = 0.8$, checkpoints `M-8` through `M-12` form the high-risk subset in Table 9. The resulting bad-anchor scale is 0.371, compared with 0.378 for heterogeneous calibration, an approximately 1.9% difference. Table 10 compares the anchors, slopes, and mappings. The bad-anchor combined score means $s_{\mathrm{ref},a} - b_a$; the slope is α_a from Equation (8).

Table 10: Heterogeneous versus reference-only calibration for `Llama`. Mapping differences are evaluated across the progressively modified checkpoints. Differences in the last column are absolute.

Quantity	Heterogeneous	Reference-only	Difference
Bad-anchor scale b_a	0.378	0.371	0.007
Bad-anchor combined score	-0.140	-0.133	0.007
Severity slope α_a	0.585	0.733	0.148
Mean absolute `RAS` difference		3.18 points	
Maximum absolute `RAS` difference		4.83 points	

The two mappings are close along the controlled intervention trajectory: their mean absolute difference is 3.18 points and their maximum difference is 4.83 points. This result supports reference-derived checkpoints as a calibration option when same-family variants are scarce. Each checkpoint contributes both its alignment score and its measured ASR to the calibration.

G Prompt-Set Robustness

The main experiment reuses prompts across direction construction, layer inspection, and alignment scoring. To examine this dependence, we use five configurations $P_i = (\mathcal{S}_i, \mathcal{U}_i, \mathcal{J}_i)$, indexed by $i \in \{1, \ldots, 5\}$. Configuration P_1 is the main configuration, and P_2–P_5 are independently constructed alternatives. Each contains 11 benign, 11 unsafe, and 11 jailbreak prompts; each jailbreak prompt extends its corresponding unsafe request with adversarial framing. Every experiment in this appendix uses `meta-llama/Llama-3.1-8B-Instruct`, the last-input-token `resid_post` activation, and, unless stated otherwise, the fixed layers 22–30 window. The four analyses examine direction stability, cross-prompt transfer, layer profiles, and prompt-count variation for this reference.

G.1 REFUSAL-DIRECTION STABILITY

For configuration P_i, let $\mu^{(i)}_{l,\text{safe}}$ and $\mu^{(i)}_{l,\text{unsafe}}$ be the layer-l reference activation means on $\mathcal{S}_i$ and $\mathcal{U}_i$. Define

$$r^{(i)}_l = \mu^{(i)}_{l,\text{unsafe}} - \mu^{(i)}_{l,\text{safe}}, \qquad \hat{r}^{(i)}_l = r^{(i)}_l / \|r^{(i)}_l\|_2, \qquad C^{(i,j)}_l = \cos(\hat{r}^{(i)}_l, \hat{r}^{(j)}_l), \tag{15}$$

where j indexes a second configuration and $C^{(i,j)}_l$ is their direction similarity. These directions are nonzero; normalization without the stabilizer gives the same cosine scores. The five configurations yield ten distinct pairs. Averaging over the scoring window, the ten pairwise similarities have mean 0.9390, standard deviation 0.0368, minimum 0.8879, and maximum 0.9740. Figure 5 displays the pairwise matrix. The high pairwise similarities show a consistent reference direction across these prompt configurations.

G.2 CROSS-PROMPT TRANSFER

Let $h_l(x) = h^{\text{ref}}_l(x)$ for this reference-model experiment, and write $\mathcal{W}$ for the fixed `Llama` scoring window. For a direction learned from P_i and evaluation prompts from P_j, define

$$u(i \to j) = \frac{1}{|\mathcal{W}||\mathcal{U}_j|} \sum_{l \in \mathcal{W}} \sum_{x \in \mathcal{U}_j} \cos(h_l(x), \hat{r}^{(i)}_l), \tag{16}$$

$$v(i \to j) = \frac{1}{|\mathcal{W}||\mathcal{J}_j|} \sum_{l \in \mathcal{W}} \sum_{x \in \mathcal{J}_j} \cos(h_l(x), \hat{r}^{(i)}_l), \tag{17}$$

$$s(i \to j) = 0.5u(i \to j) + 0.5v(i \to j). \tag{18}$$

Here $u(i \to j)$, $v(i \to j)$, and $s(i \to j)$ are the unsafe, jailbreak, and combined transfer scores. Diagonal entries are same-set evaluations; off-diagonal entries are cross-set evaluations. Their reported means are shown in Table 11. Figure 6 shows the combined-score matrix.

Table 11: Same-set versus cross-set reference alignment. Differences are same-set minus cross-set means.

Alignment	Same set	Cross set	Difference
Unsafe	0.3506	0.3342	0.0164
Jailbreak	0.2463	0.2400	0.0063
Combined	0.2984	0.2871	0.0113

Individual cross-set entries can exceed the corresponding same-set entries, while the average favors same-set scoring by 0.0113. The result indicates similar alignment across the five configurations together with a small mean dependence on construction/scoring prompt reuse.

G.3 LAYER-WINDOW ROBUSTNESS

For each configuration, define the per-layer means $m^{(i)}_{l,\text{safe}}$ and $m^{(i)}_{l,\text{unsafe}}$, suppression $q^{(i)}_l$, and gap $g^{(i)}_l$:

$$m^{(i)}_{l,\text{safe}} = \frac{1}{|\mathcal{S}_i|} \sum_{x \in \mathcal{S}_i} \cos(h_l(x), \hat{r}^{(i)}_l), \qquad m^{(i)}_{l,\text{unsafe}} = \frac{1}{|\mathcal{U}_i|} \sum_{x \in \mathcal{U}_i} \cos(h_l(x), \hat{r}^{(i)}_l),$$
$$q^{(i)}_l = -m^{(i)}_{l,\text{safe}}, \qquad g^{(i)}_l = m^{(i)}_{l,\text{unsafe}} - m^{(i)}_{l,\text{safe}}. \tag{19}$$

Figures 15–17 show similar qualitative trajectories across configurations. Unsafe alignment rises from early to middle layers and remains elevated later; suppression rises in the middle layers and stays high. The resulting gap has a broad middle-to-late high-separation region containing layers 22–30 for all five configurations. These profiles support the stability of the selected region under the tested prompt changes.

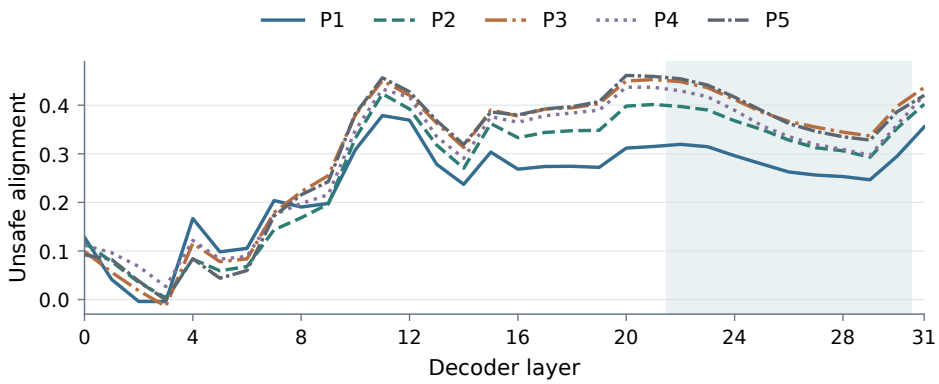


Figure 15: Unsafe-alignment profiles across five prompt configurations.

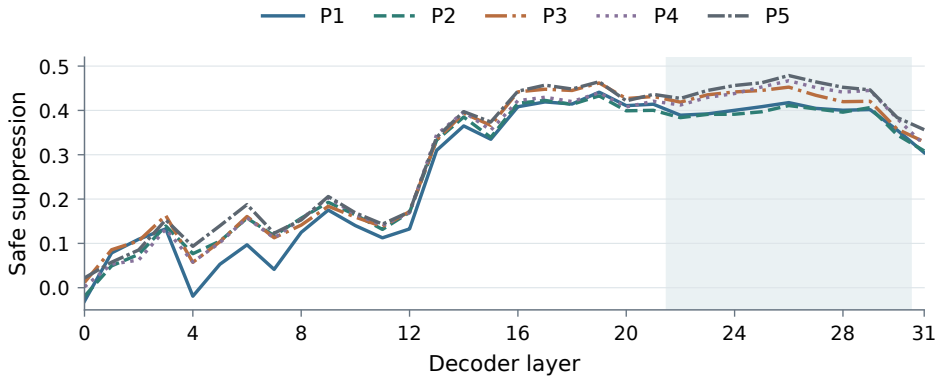


Figure 16: Safe-suppression profiles across five prompt configurations.

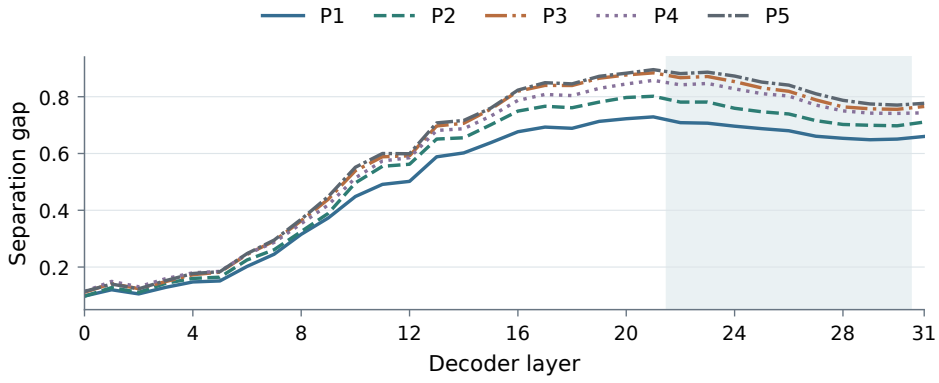


Figure 17: Separation-gap profiles across five prompt configurations.

G.4 Prompt-count convergence

We construct larger pools of 200 safe, 200 unsafe, and 200 jailbreak prompts, keep the directions learned from P_1 fixed, and score over layers 22–30. For an input x, let its mean alignment be

$$A(x) = \frac{1}{|\mathcal{W}|} \sum_{l \in \mathcal{W}} \cos(h_l(x), \hat{r}_l). \tag{20}$$

For each sample size $n \in \{1, 3, 5, 10, 11, 50, 100\}$ and each prompt type, sample n prompts without replacement and average $A(x)$. Repeat this procedure 1,000 times. Figure 18 shows the mean across resamples and empirical 95% resampling intervals; the dashed vertical line identifies $n = 11$. The intervals summarize sampling variation in the reference-model alignment estimate.

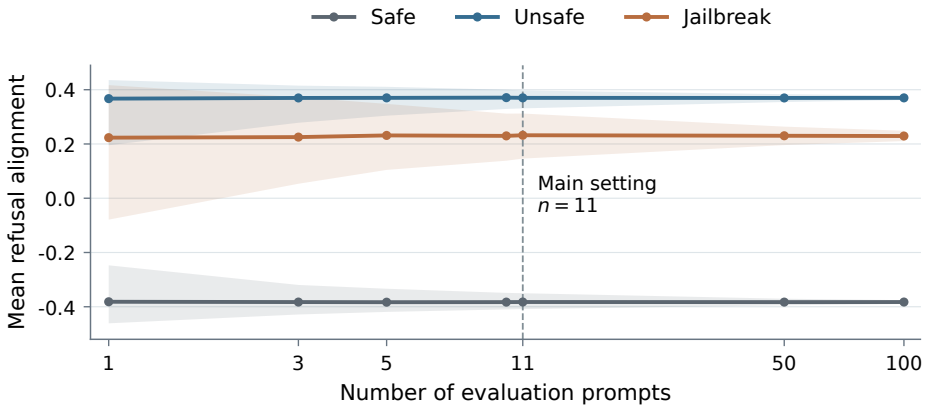


Figure 18: Prompt-count convergence under a fixed reference direction and layer window. Shaded regions are empirical resampling intervals.

Table 12: Prompt-count summaries. SD denotes standard deviation of the resampled mean. Pool means use all 200 prompts of each type.

Type	Full-pool mean	Mean at $n = 11$	SD at $n = 11$	SD at $n = 1$
Safe	−0.3826	−0.3825	0.0150	0.0528
Unsafe	0.3699	0.3700	0.0177	0.0621
Jailbreak	0.2294	0.2321	0.0415	0.1470

Table 12 shows that increasing the sample size from one to 11 reduces the standard deviation of the mean by approximately 72% for each prompt type. Jailbreak prompts remain the most variable: their standard deviation at $n = 11$ is 0.0415, compared with 0.0150 for safe and 0.0177 for unsafe prompts. Larger prompt sets further narrow the intervals in Figure 18. This supports increasing the prompt budget when precise jailbreak-alignment estimates are needed.

H Algorithm and Notation

Algorithm 1 summarizes family setup and target scoring. Manual layer selection precedes numerical calibration. Table 13 collects the notation used in the procedure.

Algorithm 1 `SafeVec` setup and target scoring for `RAS`

Require: Family set $\mathcal{A}$; references $M_{\mathrm{ref},a}$ and calibration sets $\mathcal{C}_a$ with ASR labels; prompt sets $\mathcal{S}, \mathcal{U}, \mathcal{J}$; fixed hyperparameters.

for each $a \in \mathcal{A}$ **do**
  Extract reference activations and directions $\hat{r}_l$ using Equations (1)–(2).
  Inspect reference profiles and calibration-model curves; manually choose $\mathcal{W}_a$ and freeze it before scoring targets.
  Compute $u_a(M), j_a(M), s_a(M)$ for the reference and each calibration model.
  Compute $d_a(M)$, $\mathcal{H}_a$, and $\mathcal{D}_a^+$; estimate b_a by Equation (7).
  Compute $\mathcal{V}_a$ and α_a by Equation (8); reject an undefined calibration rather than returning a safety score.
end for
Compute $\alpha_\star$ and all γ_a; freeze the family set and every scoring/calibration parameter.
for each new target M in a calibrated family a **do**
  Extract last-input-token activations on $\mathcal{U}$ and $\mathcal{J}$; compute $s_a(M)$.
  Compute $z_a(M) = \gamma_a[s_{\mathrm{ref},a} - s_a(M)]/b_a$.
  Return $\mathtt{RAS}_a(M) = 100/(1 + \exp[\beta(z_a(M) - c)])$ without generating target answers or acquiring target ASR.
end for

Table 13: Notation summary for family setup, calibration, and target scoring.

Symbol	Meaning
$\mathcal{A}, a$	Fixed calibrated-family set and a family index.
$M, M_{\mathrm{ref},a}, \mathcal{C}_a$	Evaluated model, family reference, and calibration-model set.
$\mathcal{S}, \mathcal{U}, \mathcal{J}$	Benign, direct-harmful, and jailbreak-framed input sets.
$L_a, l, D_a, h_l^M(x)$	Layer count, layer index, residual width, and last-input-token post-layer activation.
$\mu_l^{\mathrm{safe}}, \mu_l^{\mathrm{unsafe}}, r_l, \hat{r}_l$	Reference activation means, their difference, and normalized direction.
$m_l^{\mathrm{safe}}, m_l^{\mathrm{unsafe}}, q_l, g_l$	Reference alignment means, safe suppression, and separation gap.
$\mathcal{W}_a, u_a(M), j_a(M), s_a(M)$	Selected layers, two alignment means, and their weighted score.
$w_u, w_j, s_{\mathrm{ref},a}, d_a(M)$	Mixture weights, reference raw score, and reference-relative drop.
$\tau, \mathcal{H}_a, \mathcal{D}_a^+, Q_\eta, b_a$	Risk threshold, high-risk models, positive drops, empirical quantile, and anchor scale.
$\tilde{d}_a(M), \Delta_a(M), \mathcal{V}_a$	Normalized drop, ASR increase, and models eligible for slope estimation.
$\kappa_a(M), \alpha_a, \alpha_\star, \gamma_a$	Severity ratio, family slope, global median slope, and family multiplier.
$\lambda, z_a(M), c, \beta$	Severity shrinkage, aligned drop, sigmoid center, and steepness.
$\mathrm{ASR}(M), B(M), r, \rho$	Behavioral success rate, its 0–100 complement, Pearson and Spearman correlation.

I Analytical Properties

I.1 Proof of Proposition 1

Fix an admissible family calibration and abbreviate $s = s_a(M)$ and $E(s) = \exp\{\beta[\gamma_a(s_{\mathrm{ref},a} - s)/b_a - c]\}$. Both $E(s)$ and all three constants β, γ_a, b_a are positive. Differentiation gives

$$\frac{\mathrm{d}\,\mathtt{RAS}_a}{\mathrm{d}s} = \frac{100\beta\gamma_a E(s)}{b_a[1 + E(s)]^2} > 0. \tag{21}$$

Consequently, strict orderings and ties in s_a are preserved. Substituting $s = s_{\mathrm{ref},a}$ sets the alignment drop to zero and yields $\mathtt{RAS}_a(M_{\mathrm{ref},a}) = 100/(1 + \exp[-\beta c])$. With $\beta = 5$ and $c = 0.75$, this is approximately 97.70. No ASR value enters that substitution. □

I.2 Ordering and score sensitivity

Within a family, raw-score and RAS rank statistics must coincide when computed on the same models with the same tie convention. This follows directly from strict monotonicity. Across families, different reference offsets, anchor scales, and multipliers may change the ordering. Changing a common c or positive β leaves the ordering of fixed aligned drops unchanged even across families. Behavioral agreement is determined by how closely the underlying representation feature tracks harmful-request refusal.

The derivative also clarifies sensitivity. A small anchor scale increases the conversion of a raw-score change into a score change near the sigmoid center, while the sigmoid saturates at large positive or negative aligned drops. Consequently, changes in raw alignment can be amplified near the center and compressed in saturated regions. Interpreting score changes therefore requires the anchor scale and operating region as well as the underlying behavioral comparison.

Integer break-even and non-amortizing cases. When $T_{\mathrm{behavior}} > T_{\mathrm{target}}$, the smallest positive integer target count giving a strict time advantage is $N_{\min} = \max\{1, \lfloor (T_{\mathrm{setup}} + T_{\mathrm{calASR}})/(T_{\mathrm{behavior}} - T_{\mathrm{target}}) \rfloor + 1\}$, where $\lfloor \cdot \rfloor$ rounds down. If the behavioral cost is no greater than the target-scoring cost, increasing the checkpoint count cannot amortize positive setup cost into a time advantage. The cold-start bound includes the cost of acquiring calibration ASR.

I.3 Calibration arithmetic for the numerical illustration

Applying Equations (6)–(9) to the displayed reference/calibration rows gives the values in Table 14. These are calculations from the tabulated inputs, whose precision determines the last reported digits. The global median slope is $\alpha_\star \approx 0.621537$; with $\lambda = 0.5$, the `Llama` multiplier is 0.970640.

Table 14: Calibration quantities. They provide the arithmetic used in Section 3.3.

Family	$s_{\mathrm{ref},a}$	b_a	α_a	γ_a
Llama	0.238000	0.378000	0.585040	0.970640
Gemma	0.009000	0.140750	0.955620	1.268755
Qwen	0.241500	0.331000	0.621537	1.000000

Using the six-decimal checkpoint scores gives $d_a(M_{-8}) = 0.3544565$. Combining that drop with the tabulated heterogeneous-model anchor and multiplier yields $\tilde{d}_a(M_{-8}) \approx 0.937716$, $z_a(M_{-8}) \approx 0.910184$, and RAS approximately 30.98. The rounded calculation in Equation (11) reports this as about 31 points. The supplementary script computes these intermediate values directly from the two input tables.