LLM Rheology: Auditing Refusal Geometry in Aligned Language Models

Ziyang Zhao*
Huazhong University of Science and Technology
zhaoziyang@hust.edu.cn

Junlang Tian
China University of Geosciences
chris@cug.edu.cn

Abstract

Scaling improves language-model capability, but behavioral refusal does not necessarily reveal how strongly safety constraints remain coupled to adversarial task execution inside model representations. We introduce **LLM Rheology**, a representation-level framework for auditing how aligned language models respond to controlled perturbation along learned refusal directions in activation space. We define **manifold sensitivity / compliance** as the local distributional response induced by refusal-direction perturbation, measured using a finite-difference KL response with a local Fisher-information interpretation. This quantity provides a representation-level measure of task–refusal coupling under adversarial stress. Across six open model families and 22 checkpoints, including Qwen, DeepSeek-Distill, Mistral, Llama-3, Gemma, and Yi, we observe recurring, but family-dependent, scale-structured variation. Smaller aligned models often exhibit elevated refusal response, intermediate-scale checkpoints frequently enter weakened-response regimes, and several larger checkpoints recover toward near-baseline response. Cosine measurements on the Qwen scaling axes further show decreasing alignment between refusal directions and task-generation subspaces as scale increases. We additionally provide causal evidence through inference-time activation intervention on Qwen-72B. Injecting a learned refusal vector at a critical semantic layer restores refusal behavior on the evaluated jailbreak subset while largely preserving benign reasoning performance, whereas matched-norm placebo vectors fail to reproduce the effect. Together, these findings suggest that representation-level task–refusal coupling can evolve non-monotonically with scale, even when surface-level refusal behavior remains strong. LLM Rheology provides a complementary geometric perspective for studying refusal safety beyond behavioral evaluation.

1 Introduction

Large language models exhibit robust empirical scaling in capability: increasing data, parameters, and compute improves loss, reasoning, and downstream performance in predictable ways [22, 17]. Safety, however, need not scale by the same mechanism. Instruction tuning and RLHF-style training can improve observed helpfulness and harmlessness [28, 6], but these alignment techniques often yield demographic-specific safeguards rather than universal constraints [8]. Consequently, behavioral refusal alone does not reveal whether a model has internalized a robust safety constraint, merely learned a policy-level filter, or geometrically separated refusal features from task execution.

This distinction matters because modern LLM failures increasingly appear as *latent* rather than purely surface-level failures. Jailbreaks show that aligned models can be pushed into harmful-compliance regimes despite apparently safe behavior under ordinary prompts [37, 43], with vulnerability scaling exponentially under adversarial inference optimization [15, 36]. Backdoor and sleeper-agent studies

*Corresponding author.

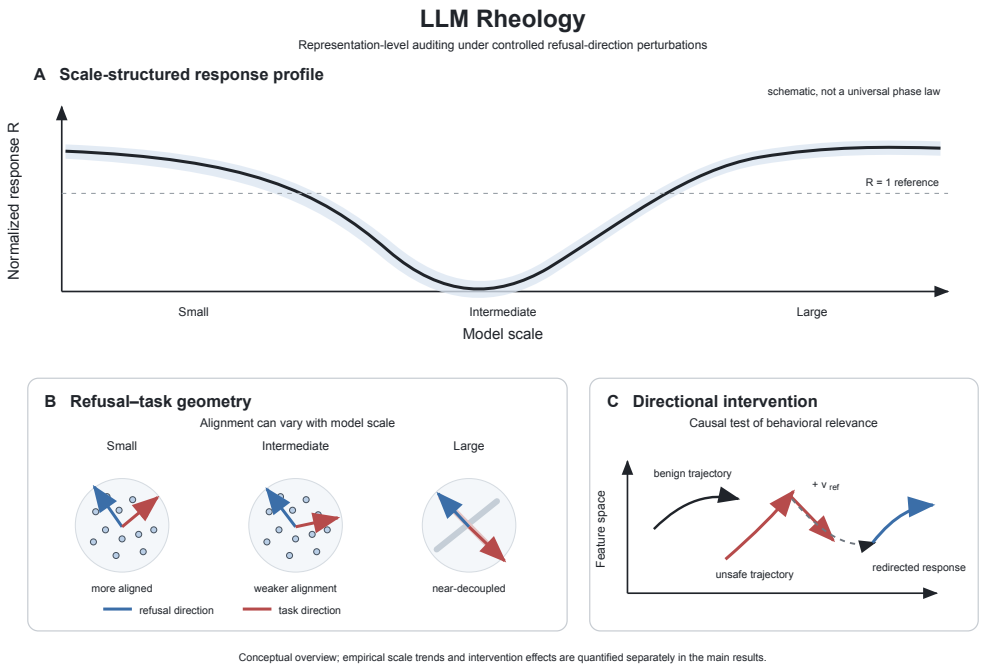


Figure 1: **Overview of LLM Rheology.** We probe refusal geometry by applying controlled perturbations along an empirically estimated refusal direction and measuring the resulting distributional response. **Top:** Schematic illustration of a possible scale-structured response profile. The curve provides conceptual motivation and does not posit a universal three-stage scaling law across model families. **Bottom left:** Geometric interpretation in which the alignment between refusal and task-generation directions can vary with model scale. **Bottom right:** Direction-specific activation intervention provides a causal test of whether the learned refusal representation can alter harmful-content behavior while retaining benign capability.

further show that undesirable behaviors can persist after safety training [18]. Recent work also suggests that deceptive behavior need not always be explicitly prompt-induced: Wu et al. [40] report that LLMs can exhibit deception-like inconsistency even under benign prompts, exacerbating as task difficulty rises. These findings motivate a more structural question: *what happens to the geometry of safety constraints as models scale?*

We propose **LLM Rheology**, a representation-level framework for auditing how aligned language models respond to adversarial stress inside activation space. The analogy to physical rheology is diagnostic rather than literal: instead of asking only whether a model refuses, we measure how its output distribution changes when a controlled perturbation is applied along a learned refusal direction.

Concretely, let v_{ref} denote a refusal direction extracted from behaviorally filtered contrastive pairs. For an activation $h^{(\ell)}$ at layer ℓ, we apply

$$h^{(\ell)} \leftarrow h^{(\ell)} + \alpha v_{\mathrm{ref}},$$

and measure the induced distributional response using a local KL/Fisher–Rao approximation. We refer to the resulting scalar as **manifold sensitivity / compliance**: a representation-level measure of how strongly refusal-related features remain coupled to adversarial task-generation trajectories.

Across six open model families and 22 checkpoints, we observe family-dependent, scale-structured variation rather than monotonic safety improvement. Smaller aligned models often show elevated response to refusal-direction perturbations, intermediate-scale checkpoints frequently show weakened response, and several larger checkpoints recover toward a near-baseline regime. We treat these patterns as phenomenological observations rather than evidence for a universal scaling law.

We support this interpretation with three pieces of evidence. First, normalized response ratios exhibit similar non-monotonic structure across multiple model families. Second, cosine measurements on

the Qwen scaling axes show decreasing alignment between refusal directions and task-generation subspaces as scale increases. Third, a localized activation intervention on Qwen-72B restores refusal behavior on the evaluated jailbreak subset while largely preserving benign reasoning performance, suggesting that the learned refusal direction remains causally meaningful even when its coupling to adversarial trajectories is weak.

Our contributions are:

1. We introduce **LLM Rheology**, a representation-level auditing framework for measuring refusal safety through activation-space stress–response behavior.
2. We define **manifold sensitivity / compliance** as a KL/Fisher–Rao-normalized response metric and apply it across six model families and 22 checkpoints.
3. We identify a recurring scale-structured response pattern consisting of elevated small-model response, intermediate weakened-response regimes, and large-model recovery toward near-baseline geometry.
4. We provide supporting mechanistic and causal evidence for task–refusal decoupling through cosine analysis and localized activation intervention on Qwen-72B.

2 Related Work

Our work connects scaling laws, alignment failures, deception, and representation-level interventions, treating refusal robustness as a *geometric response property* under stress rather than a mere binary output.

Scaling laws and safety boundaries. While capability scaling improves predictability [22, 17], behaviors often exhibit non-monotonic or inverse-scaling trends [26, 38]. In safety, RLHF improves surface alignment [28, 6], but scale can exacerbate sycophancy, latent risks [30, 39, 9], and demographic safety disparities [8]. Recent Bayesian scaling models further reveal how post-training fails to prevent harmful behaviors under in-context learning [4]. We add a representation-level account: safety decays geometrically when refusal features decouple from task generation.

Jailbreaks, deception, and latent failures. Safety training is frequently bypassed by adversarial jailbreaks [37, 43, 21], often because models imitate preferred behavior without internalizing robust constraints [10]. This latent vulnerability allows deceptive policies [18] and prompt-induced lies [40] to persist. Recent dissections of LLM refusal demonstrate that these safety failures stem from exploitable internal representations [31]. We address this by directly measuring task–refusal coupling in activation space.

Mechanistic interpretability and linear geometry. Moving beyond isolated neurons, recent work emphasizes distributed features, the linear representation hypothesis [29], superposition [13], and dictionary learning [7]. Building on these, we shift focus from static feature discovery to *dynamic response*, measuring how the representation space deforms when refusal directions are stressed.

Representation engineering and interventions. Activation steering and concept erasure edit model behaviors via linear activation-space operations [44, 34, 19, 33]. However, recent findings show that steering techniques can act as a "rogue scalpel," inadvertently compromising safety safeguards [23], necessitating new methods like representation bending to disrupt harmful generation [42]. Unlike global concept erasure, our diagnostic intervention tests whether surgical, localized refusal-vector injection can restore semantic-control barriers without degrading benign computation.

Trajectory-level latent safety diagnostics. Recent work such as N-GLARE [24] and Semantic Turbulence [32] studies safety through passive latent-trajectory separability or within-trajectory instability. LLM Rheology is complementary: rather than measuring whether trajectories separate under ordinary inference, we apply an active local stress along a learned refusal direction and measure the resulting distributional response. Related latent-space attack and defense methods, including LatentBreak [27] and LATPC [41], also emphasize that safety can be mediated by representation geometry; our contribution is to audit how refusal-task coupling changes with scale and layer.

3 LLM Rheology: A Framework for Safety Response Geometry

We introduce **LLM Rheology**, a representation-level framework for measuring how aligned language models respond to adversarial perturbation inside activation space. Rather than evaluating safety only through surface behavior (*e.g.*, refusal strings), we measure the local distributional response induced by perturbation along a learned refusal direction. Our framework assumes distributed dense representations in high-dimensional activation space, where semantically independent concepts are approximately quasi-orthogonal due to concentration-of-measure effects [35]. This approximation is expected to hold primarily in sufficiently over-parameterized Transformer regimes, and may become weaker in very small or highly sparse models. The central quantity we study is **manifold sensitivity / compliance**, which operationalizes how strongly refusal behavior remains coupled to task-generation dynamics in representation space.

3.1 Representation Geometry

Let $\mathcal{M}_\theta \colon \mathcal{X} \to \Delta(\mathcal{Y})$ denote a language model parameterized by θ, mapping an input sequence $x \in \mathcal{X}$ to a probability distribution over vocabulary $\mathcal{Y}$. We analyze hidden activations at layer ℓ, denoted $\mathcal{H}^{(\ell)} \subseteq \mathbb{R}^d$, with hidden state

$$h^{(\ell)} = \mathcal{M}_\theta^{(1:\ell)}(x).$$

Following prior work on linear representation geometry and feature superposition [29, 13, 44], semantic properties such as refusal and task-generation behavior can often be approximately represented as directions or low-dimensional subspaces in activation space [14].

3.2 Stress and Distributional Response

We define *semantic stress* as a directed perturbation along a learned refusal direction $v_{\text{ref}} \in \mathbb{R}^d$ [3]. The perturbation is:

$$\delta = \alpha v_{\text{ref}}, \qquad \sigma = \|\delta\|_2, \tag{1}$$

where α controls perturbation magnitude.

The refusal direction is extracted from behaviorally filtered contrastive prompt pairs (Appendix B).

To measure the induced response, we compute the local deformation of the model's next-token distribution using KL divergence:

$$\varepsilon(\delta) = \sqrt{D_{\text{KL}}\left(\mathcal{M}_\theta(\cdot \mid h^{(\ell)}) \parallel \mathcal{M}_\theta(\cdot \mid h^{(\ell)} + \delta)\right)}. \tag{2}$$

Under the local Fisher approximation [2],

$$D_{\text{KL}} = \frac{\alpha^2}{2} v_{\text{ref}}^\top I(h) v_{\text{ref}} + o(\alpha^2), \tag{3}$$

where $\mathcal{I}(h)$ is the Fisher Information Matrix. The square root therefore yields a locally linear response magnitude comparable to the perturbation norm.

3.3 Manifold Sensitivity / Compliance

Definition 3.1 (Manifold Sensitivity / Compliance)**.** Let v_{ref} be the unit refusal vector and σ, ε be as defined above. The *manifold sensitivity / compliance* is:

$$k(x, v_{\text{ref}}; \alpha) \triangleq \frac{D_{\text{KL}}(p_h \parallel p_{h+\alpha v_{\text{ref}}})}{\alpha^2 \|v_{\text{ref}}\|_2^2} \to \frac{1}{2} \frac{v_{\text{ref}}^\top I(h) v_{\text{ref}}}{\|v_{\text{ref}}\|_2^2}., \tag{4}$$

measured at the critical semantic-control layer ℓ^* (see Section 4.5).We therefore treat k_{FD} as the directly measured finite-difference response, while its Fisher interpretation is restricted to the empirically verified local perturbation regime.An explicit perturbation-scale sweep confirms an approximately quadratic KL regime over $|\alpha| = 0.0025 - 0.01$ (fitted log–log slope 1.96; Appendix I.2)

Remark 3.2 (Relationship to classical stiffness). Classical stiffness is usually *stress/strain*, measuring resistance to deformation. Our quantity k is instead the response ratio *strain/stress*. We use this convention because, in the alignment setting, a larger distributional response per unit refusal-direction injection indicates that the learned refusal feature remains active and coupled to generation. Conventional mechanical rigidity therefore corresponds to $1/k$, whereas high k here means high *safety-response sensitivity*. Locally, this quantity is proportional to the Fisher–Rao norm of the refusal feature [2]. Our Fisher–Rao interpretation is local and therefore depends on small perturbation regimes. In the present work we use a fixed perturbation scale ($\delta_0 = 0.01$) and a consistent final-instruction-token measurement protocol across families.

Normalized response ratio (R). To isolate safety-specific response from baseline generation geometry, we compare adversarial prompts against matched controls:

$$R = \frac{k_{\text{treat}}}{k_{\text{control}}}, \tag{5}$$

where k_{treat} is measured on adversarial jailbreak prompts and k_{control} on behaviorally matched non-jailbreak controls. The reference point $R = 1$ corresponds to equal response between adversarial and matched-control prompts, which we treat as the empirical baseline regime. Values $R > 1$ indicate elevated refusal response under adversarial stress, while $R < 1$ indicates weakened response. Near-baseline response ($R \approx 1$) indicates that adversarial prompts no longer induce substantially stronger geometric response than their matched controls.

3.4 Mechanism: Geometric Orthogonalization Hypothesis

Why does the aligned-model safety response tend to decay toward a near-isotropic baseline as scale increases? Let $\mathcal{S} \subset \mathbb{R}^d$ denote the refusal/safety subspace, and let $\mathcal{C} \subset \mathbb{R}^d$ denote the task-generation subspace, including adversarial or harmful-compliance directions. By concentration of measure in high-dimensional spaces [35], independent directions in $\mathbb{R}^d$ become nearly orthogonal with high probability as $d \to \infty$:

$$\cos\theta_{\mathcal{S},\mathcal{C}} = \frac{\langle v_{\text{safe}}, v_{\text{task}} \rangle}{\|v_{\text{safe}}\| \, \|v_{\text{task}}\|} \;\longrightarrow\; 0 \quad \text{as } d \to \infty. \tag{6}$$

We hypothesize that large-scale safety tuning increasingly exploits this geometric slack: rather than eliminating harmful task-generation features, it can separate refusal and task-generation features into weakly coupled directions [5]. Under such *orthogonal relaxation*, adversarial task execution can proceed with little projection onto the refusal subspace, so perturbations along v_{ref} induce smaller distributional response. This provides a geometric account of the scale-structured regime transition in R and the extreme-scale normalization regime analyzed in Section 4. We provide a high-dimensional geometric derivation and toy-model motivation in Appendix A.

3.5 Feature Crowding and Geometric Frustration

The weakened responses observed at several intermediate checkpoints may reflect that refusal weakening is not purely monotonic, but may reflect finite-dimensional interference. As scale increases, reasoning and task-execution features appear to grow faster than the hidden dimension available for approximately orthogonal allocation. Safety-related and task-related directions may therefore partially entangle, producing destructive gradient interference between helpfulness and refusal objectives [5].

We use geometric frustration as a descriptive hypothesis for this possible finite-dimensional interference. Empirically, Qwen-2.5 shows maximal response collapse in the 14B–32B regime before partially recovering at 72B, while task–refusal cosine similarity monotonically approaches the high-dimensional orthogonality limit. This pattern is consistent with temporary feature crowding followed by large-scale orthogonal relaxation. Additional quantitative analysis is provided in Appendix H.

4 Experiments

4.1 Experimental Setup

Models. We evaluate 22 checkpoints across six open model families: Qwen-2.5, DeepSeek-R1-Distill, Mistral, Llama-3, Gemma-2, and Yi, covering scales from 3B to 405B parameters. Experi-

ments are conducted using behaviorally matched treat/control prompt pairs to measure rheological response under directed refusal perturbation. Across eight paired Qwen and DeepSeek-R1-Distill checkpoints, BF16 and Int4 measurements are highly correlated ($r > 0.98$) and preserve the principal within-series trends and critical-layer locations; paired absolute differences are reported in Appendix G.

Datasets. Adversarial prompts are drawn from JailbreakBench [11], AdvBench [43], and HarmBench. Benign capability controls use GSM8K [12] and MMLU [16] subsets. All reported confidence intervals are 95% bootstrap confidence intervals computed over prompt subsets with 1000 resamples.

Refusal vector extraction. We extract a refusal direction from behaviorally filtered contrastive prompt pairs, where one prompt induces refusal and the matched counterpart does not. At the critical layer ℓ^*, the refusal vector is computed as the normalized mean activation difference:

$$\tilde{v}_{\text{ref}} = \frac{1}{N}\sum_{i=1}^{N}\left(h^{(\ell^*)}(x_i^+) - h^{(\ell^*)}(x_i^-)\right), \qquad v_{\text{ref}} = \frac{\tilde{v}_{\text{ref}}}{\|\tilde{v}_{\text{ref}}\|_2}. \tag{7}$$

Implementation. Experiments are conducted on NVIDIA H800 GPUs using `vLLM` and PyTorch. Within each family, refusal and task directions are extracted using the native chat template and system-prompt conventions of that family. Because all comparisons are performed within-family before cross-family aggregation, the reported trends are less likely to be explained purely by prompt-format differences. Nevertheless, prompt-template effects remain a potential source of representational variation and warrant further controlled study. Additional implementation details, dataset statistics, filtering procedures, and hyperparameters are provided in Appendix B and G.

4.2 Scale-Structured Variation in Normalized Response

Figure 2 and Table 1 summarize the normalized response ratio R across all six families. The measurements show non-monotonic scale-associated variation rather than a universal monotonic trend. Because checkpoint coverage differs substantially across families, we do not interpret these observations as a universal three-phase scaling law. Instead, we distinguish the two densely sampled series from the remaining sparse cross-family reference points.

For Qwen-2.5, we summarize the observed middle-checkpoint reduction using the descriptive contrast

$$\Delta_{\text{Qwen}} = (R_{7B} + R_{72B})/2 - (R_{14B} + R_{32B})/2$$

yielding 0.255 with a paired-bootstrap 95% CI of $[0.078, 0.432]$. For DeepSeek-R1-Distill, whose sampled series contains a single pronounced minimum at 32B, we use

$$\Delta_{\text{DeepSeek}} = (R_{7B} + R_{14B} + R_{70B})/3 - R_{32B}$$

yielding 0.443 with a 95% CI of $[0.195, 0.691]$. These contrasts are descriptive summaries of the observed series rather than a common pre-specified test. The DeepSeek-R1-Distill series is not a backbone-controlled scaling axis because its 70B endpoint is Llama-based.

As an exploratory convergent-validity analysis, we additionally compare R with independently evaluated harmful-content ASR across the eight checkpoints in the two densely sampled series, obtaining checkpoint-level Spearman $\rho = -0.64$ in Appendix I.4. Given the small number of checkpoint-level units, we do not interpret this as definitive predictive validation.

4.3 Microscopic Mechanism: Asymptotic Orthogonality

For cosine analysis, we estimate the task-generation direction v_{task} using the normalized mean activation difference between harmful-compliance continuations and matched benign-control continuations at the same critical layer ℓ^* used for refusal-vector extraction. Unless otherwise stated, activations are taken at the final instruction token before generation. We use a single contrastive direction rather than a PCA subspace in the main analysis, leaving multi-directional task subspaces to future work. Cosine similarity is then computed as $\cos(v_{\text{ref}}, v_{\text{task}})$. Table 2 reports cosine similarity values for

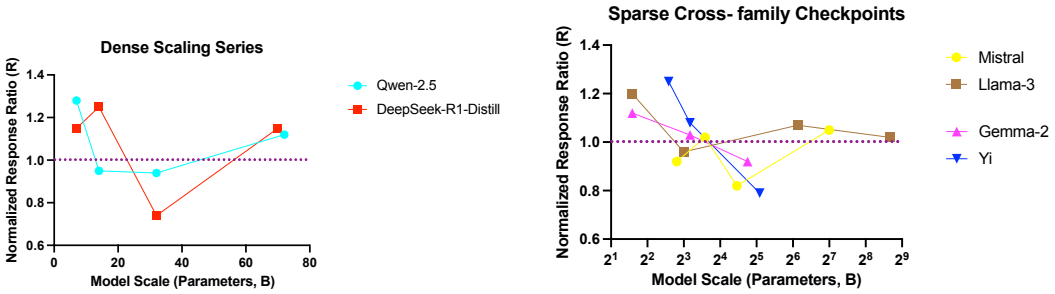


Figure 2: **Scale-structured variation of normalized response.** The normalized response ratio $R = k_{\text{treat}}/k_{\text{control}}$ varies non-monotonically with model scale. Qwen-2.5 and DeepSeek-R1-Distill provide the densest sampled series, while the remaining families serve as sparse cross-family reference points. The dashed line $R = 1$ denotes equal response between adversarial and matched-control prompts. We do not interpret the observed structure as a universal three-phase scaling law.

Table 1: **Normalized Response Ratio R Across Six Families.** $R = k_{\text{treat}}/k_{\text{control}}$. $R = 1$ denotes equal normalized response between adversarial and matched-control prompts. Values represent means over N=500 matched prompt pairs. The final column indicates whether the paired-bootstrap 95% confidence interval lies entirely above 1, entirely below 1, or overlaps 1. Overlap with 1 is not interpreted as statistical equivalence. Dagger (†) denotes distillation-based checkpoints.

Family	**Size**	R	**95% CI relative to $R = 1$**
Qwen-2.5	7B	1.28	entirely above 1
	14B	0.95	overlapping 1
	32B	0.94	overlapping 1
	72B	1.12	entirely above 1
DeepSeek-R1-Distill†	7B	1.15	entirely above 1
	14B	1.25	entirely above 1
	32B	0.74	entirely below 1
	70B	1.15	entirely above 1
Mistral	7B	0.92	entirely below 1
	12B (Nemo)	1.02	overlapping 1
	22B	0.82	entirely below 1
	128B	1.05	overlapping 1
Llama-3	3B	1.20	entirely above 1
	8B	0.96	overlapping 1
	70B	1.07	entirely above 1
	405B	1.02	overlapping 1
Gemma-2	2B	1.12	entirely above 1
	9B	1.03	overlapping 1
	27B	0.92	entirely below 1
Yi	6B	1.25	entirely above 1
	9B	1.08	entirely above 1
	34B	0.79	entirely below 1

Qwen-2.5.This supports a transition from clearly non-null refusal–task alignment at 7B toward the empirical unrelated-direction range at 72B.

4.4 Depth-Wise Analysis: Alignment as a High-Order Semantic Constraint

Depth-wise probing shows that refusal geometry is unevenly distributed across the Transformer hierarchy. Across all six evaluated families, manifold sensitivity peaks in mid-to-deep layers (approximately $50\% - 80\%$ relative depth), while shallow layers show weak response and final layers yield diminishing intervention effectiveness.

Table 2: **Refusal–task cosine similarity on Qwen-2.5.** Cosine similarity decreases monotonically with scale. Relative to 500 unrelated semantic-direction pairs estimated using the same extraction pipeline, the 7B value lies above more than 99% of empirical-null values, whereas the 72B value lies near the 58th percentile.

Model	7B	14B	32B	72B
$\cos\theta(v_{\text{ref}}, v_{\text{task}})$	0.085	0.042	0.021	0.012
empirical-null percentile	99th	95th	86th	58th

This localization suggests that alignment-related representations emerge primarily in higher-order semantic regions, motivating our use of a critical semantic-control layer ℓ^*. Extended depth-wise analyses, layer-wise heatmaps, and family-specific dynamics are provided in Appendix D.

4.5 Critical Layer Selection

To compare models with different depths (28 layers for 7B vs. 80 layers for 72B), we perform a two-stage depth sweep to identify the critical safety layer ℓ^* for each model.

Stage 1: Refusal vector norm scan. We scan layers and select the depth ℓ maximizing $\|v_{\text{ref}}^{(\ell)}\|_2$, indicating where the refusal direction is most semantically concentrated. This consistently falls within Zone 2 (relative depth 0.6–0.8), matching the Semantic Control Plateau analysis.

Stage 2: Intervention trade-off validation. We test surgical interventions (Section 4.6) at candidate layers from Stage 1, selecting the layer that minimizes attack success rate (ASR) while preserving benign accuracy. Layers before Zone 2 insufficiently suppress jailbreaks; Zone 3 layers degrade benign fluency. The resulting critical layers are ℓ^*=55 for Qwen-72B (relative depth 0.69), ℓ^*=24 for Mistral-7B (0.75), and ℓ^*=32 for Gemma-2-9B (0.76).

4.6 Causal Intervention

To test whether Qwen-72B's near-baseline response reflects weak task–refusal coupling, we perform inference-time activation intervention at the critical semantic layer ($\ell^* = 55$).

Intervention protocol. We modify the hidden state during the forward pass via:

$$h_{\text{modified}}^{(\ell^*)} = h_{\text{original}}^{(\ell^*)} + \alpha v_{\text{ref}}, \tag{8}$$

where v_{ref} is the learned refusal direction and α controls intervention strength. Model weights remain unchanged. Figure 3 shows a narrow capability-preserving intervention window near $\alpha = 0.2$, where jailbreak success rate drops substantially while benign reasoning remains largely unchanged. Larger intervention strengths begin to degrade capability. ASR is measured using a rule-based refusal/compliance rubric consistent with HarmBench-style evaluation, with responses classified as successful jailbreaks if they contain actionable harmful-compliance content. We use α primarily as a local causal probing parameter, not a universally calibrated steering coefficient. The key observation is a narrow window in which refusal behavior can be restored without substantially perturbing benign reasoning. Matched-norm random placebo vectors fail to reproduce the same pattern, suggesting that the effect depends on semantic direction rather than perturbation magnitude alone. Additional analyses and layer-wise ablations are provided in Appendix E. We further repeat the intervention on Llama-3-70B using disjoint extraction, layer-selection, operating-point-selection, and final-test splits. At the selected operating point, the intervention reduces harmful-content ASR by 58.4 percentage points in Appendix I.3. We treat this as directionally consistent second-family evidence; the same capability-preserving trade-off has not been statistically established for Llama-3-70B. Deterministically selected held-out qualitative cases, including all intervention controls, are provided in Appendix I.1.

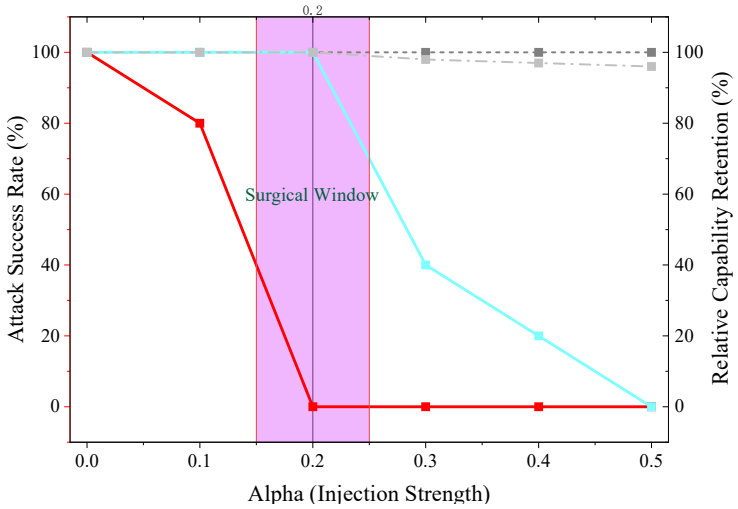


Figure 3: **Capability-preserving intervention window.** Injecting the refusal vector reduces jailbreak success rate on the evaluated subset while largely preserving GSM8K performance near $\alpha = 0.2$. Larger intervention strengths begin to degrade capability. Matched-norm placebo vectors do not reproduce the effect, supporting directional specificity of the intervention. Capability scores are reported relative to baseline performance on the evaluated GSM8K subset ($N = 500$).At the predeclared operating point $\alpha = 0.2$, refusal-vector injection reduces harmful-content ASR by 76.2 percentage points relative to baseline ($95\%CI[68.9, 83.5]$) and by 74.8 points relative to the matched-norm random control ($95\%CI[67.2, 82.4]$), while GSM8K accuracy decreases by 1.4 points. The intervention is therefore interpreted as a local causal probe rather than a generally calibrated safety mechanism.

5 Discussion and Limitations

Our results suggest that refusal safety is not only behavioral, but also geometric: larger checkpoints often exhibit weaker coupling between refusal directions and adversarial task trajectories even when surface-level refusal remains strong. One interpretation is that RLHF-style alignment may not remove harmful-compliance capabilities, but instead weakly decouples them from refusal-related directions in activation space. This view is broadly consistent with latent-failure phenomena such as sleeper agents [18], where undesirable behaviors can persist despite safety training. In several larger checkpoints, we observe weaker coupling between adversarial task trajectories and the learned refusal direction, accompanied by a return toward near-baseline response. These findings highlight limitations of purely behavioral evaluation. Benchmarks such as JailbreakBench and HarmBench [11, 25] measure refusal or compliance, but not how strongly refusal-related representations remain coupled to adversarial task execution. LLM Rheology complements such benchmarks with representation-level diagnostics based on stress–response behavior, layer localization, and task–refusal coupling.

Our study has several limitations. First, rheological measurement and intervention require white-box activation access, and quantization may still introduce subtle activation-level distortions despite matched BF16-vs-Int4 consistency checks. Second, the clearest evidence for non-monotonic scale-associated variation comes from the densely sampled Qwen-2.5 and DeepSeek-R1-Distill series, while the remaining families provide sparser cross-scale reference points and do not establish a common scaling trajectory. Third, our geometric interpretation relies on approximate high-dimensional quasi-orthogonality assumptions that may weaken in very small or sparse regimes. Fourth, the framework focuses on a single contrastive refusal direction and averaged prompt statistics rather than the full safety manifold; extending the analysis to multi-dimensional subspaces (e.g., PCA, SAEs, or dictionary features) remains important future work. Finally, the intervention experiments use locally selected perturbation scales and minimal placebo controls, serving primarily as causal probes rather than universally calibrated steering mechanisms.

Overall, our analysis should be interpreted as a geometric and phenomenological characterization of observed scaling behavior rather than a formal theory of LLM optimization dynamics. Broader validation across additional model families, architectures, training seeds, and evaluation distributions remains important future work.

6 Conclusion

We introduced **LLM Rheology**, a representation-level framework for auditing refusal safety through activation-space response geometry. Instead of asking only whether a model refuses, it measures how strongly hidden states respond to perturbations along a learned refusal direction. Across six model families and 22 checkpoints, we find a family-dependent scale-structured variation: elevated small-model response, weakened intermediate-scale regimes, and partial large-model recovery toward near-baseline response. These trends are consistent with increasing decoupling between refusal-related and task-generation directions as representation dimensionality grows, while remaining phenomenological rather than a universal scaling law.

We also provide causal evidence on Qwen-72B: localized refusal-vector injection restores refusal behavior on the evaluated jailbreak subset while largely preserving benign reasoning. Together, these results suggest that internal response geometry can complement behavioral safety evaluation by revealing how refusal representations are localized, coupled, and stressed inside large language models. More broadly, *Rheological Profiling* offers a diagnostic lens for studying safety-relevant representations beyond surface-level refusals.

Acknowledgments

The authors thank the anonymous reviewers of an earlier version of this work for their constructive comments and suggestions.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests Disclosure

The authors declare no competing interests.

References

[1] Pierre-Antoine Absil, Robert E. Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. 2007. URL https://api.semanticscholar.org/CorpusID:6769887.

[2] Shun-ichi Amari and Hiroshi Nagaoka. Methods of information geometry. 2000. URL https://api.semanticscholar.org/CorpusID:116976027.

[3] 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, 2024. URL https://arxiv.org/abs/2406.11717.

[4] Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D. Goodman. Bayesian scaling laws for in-context learning, 2025. URL https://arxiv.org/abs/2410.16531.

[5] Sadia Asif and Mohammad Mohammadi Amiri. Refusalguard: Geometry-preserving fine-tuning for safety in llms, 2026. URL https://arxiv.org/abs/2605.01913.

[6] 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, 2022. URL https://arxiv.org/abs/2212.08073.

[7] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/.

[8] Iago Alves Brito, Walcy Santos Rezende Rios, Julia Soares Dollis, Diogo Fernandes Costa Silva, and Arlindo Rodrigues Galvão Filho. Safety is not universal: The selective safety trap in llm alignment, 2026. URL https://arxiv.org/abs/2601.04389.

[9] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023. URL https://arxiv.org/abs/2312.09390.

[10] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

[11] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024. URL https://arxiv.org/abs/2310.08419.

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

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

[14] Lang Gao, Jinghui Zhang, Wei Liu, Fengxian Ji, Chenxi Wang, Zirui Song, Akash Ghosh, Youssef Mohamed, Preslav Nakov, and Xiuying Chen. The cylindrical representation hypothesis for language model steering, 2026. URL https://arxiv.org/abs/2605.01844.

[15] Indranil Halder, Annesya Banerjee, and Cengiz Pehlevan. Jailbreak scaling laws for large language models: Polynomial-exponential crossover, 2026. URL https://arxiv.org/abs/2603.11331.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

[17] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

[18] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

[19] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic, 2023. URL https://arxiv.org/abs/2212.04089.

[20] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

[21] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models, 2023. URL https://arxiv.org/abs/2309.00614.

[22] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[23] Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. The rogue scalpel: Activation steering compromises llm safety, 2026. URL https://arxiv.org/abs/2509.22067.

[24] Zheyu Lin, Jirui Yang, Yukui Qiu, Hengqi Guo, Yubing Bao, and Yao Guan. N-glare: An non-generative latent representation-efficient llm safety evaluator, 2026. URL https://arxiv.org/abs/2511.14195.

[25] 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, 2024. URL https://arxiv.org/abs/2402.04249.

[26] Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better, 2024. URL https://arxiv.org/abs/2306.09479.

[27] Raffaele Mura, Giorgio Piras, Kamilė Lukošiūtė, Maura Pintor, Amin Karbasi, and Battista Biggio. Latentbreak: Jailbreaking large language models through latent space feedback, 2025. URL https://arxiv.org/abs/2510.08604.

[28] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[29] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

[30] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

[31] Nirmalendu Prakash, Yeo Wei Jie, Amir Abdullah, Ranjan Satapathy, Erik Cambria, and Roy Ka Wei Lee. Beyond i'm sorry, i can't: Dissecting large language model refusal, 2026. URL https://arxiv.org/abs/2509.09708.

[32] Md. Hasib Ur Rahman. The laminar flow hypothesis: Detecting jailbreaks via semantic turbulence in large language models, 2025. URL https://arxiv.org/abs/2512.13741.

[33] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure, 2024. URL https://arxiv.org/abs/2201.12091.

[34] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

[35] Roman Vershynin. High-dimensional probability. 2018. URL https://api.semanticscholar.org/CorpusID:196128750.

[36] Xiangwen Wang, Ananth Balashankar, and Varun Chandrasekaran. Systematic scaling analysis of jailbreak attacks in large language models, 2026. URL https://arxiv.org/abs/2603.11149.

[37] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. URL https://arxiv.org/abs/2307.02483.

[38] Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. Inverse scaling can become U-shaped. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15580–15591, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.963. URL https://aclanthology.org/2023.emnlp-main.963/.

[39] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2024. URL https://arxiv.org/abs/2308.03958.

[40] Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. Beyond prompt-induced lies: Investigating llm deception on benign prompts, 2026. URL https://arxiv.org/abs/2508.06361.

[41] Xin Yi, Yue Li, Dongsheng Shi, Linlin Wang, Xiaoling Wang, and Liang He. Latent-space adversarial training with post-aware calibration for defending large language models against jailbreak attacks, 2025. URL https://arxiv.org/abs/2501.10639.

[42] Ashkan Yousefpour, Taeheon Kim, Ryan S. Kwon, Seungbeen Lee, Wonje Jeung, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, and Jonghyun Choi. Representation bending for large language model safety, 2025. URL https://arxiv.org/abs/2504.01550.

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

[44] 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, 2025. URL https://arxiv.org/abs/2310.01405.

A Geometric Derivation and Toy-Model Motivation for Orthogonal Relaxation

The following analysis should be interpreted as a high-dimensional geometric approximation rather than a formal proof of optimization dynamics in large language models. Our goal is to provide an interpretable mechanism consistent with the observed scaling trends, rather than to claim that trained representations behave as fully random vectors.

We provide a mathematical foundation for the Orthogonal Decoupling Hypothesis, drawing on high-dimensional probability theory [35], the Neural Tangent Kernel [20], and manifold optimization [1].

A.1 Problem Formulation

Consider a deep network $f(x;\theta)$ with activation space $\mathcal{H} \cong \mathbb{R}^d$ at layer ℓ. Define competing objectives:

$$\mathcal{L}_{\text{task}}(h) : \text{jailbreak / deceptive-task objective}, \tag{9}$$

$$\mathcal{L}_{\text{safe}}(h) : \text{safety alignment objective (refusal)}. \tag{10}$$

Under RLHF, the total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{safe}}$.

Definition A.1 (Manifold Sensitivity)**.** Following Eq. (4):

$$k \propto \frac{\partial^2 \mathcal{L}_{\text{safe}}}{\partial v_{\text{ref}}^2} \approx v_{\text{ref}}^\top H_{\text{safe}}\, v_{\text{ref}}, \tag{11}$$

where H_{safe} is the Hessian of the safety loss.

A.2 Concentration of Measure

Lemma A.2 (Quasi-Orthogonality [35])**.** *Let $u, v \in \mathbb{R}^d$ be independent uniform unit vectors. For any $t > 0$:*

$$\Pr\Big[|\langle u, v\rangle| \geq \tfrac{t}{\sqrt{d}}\Big] \leq 2e^{-ct^2}. \tag{12}$$

Theorem A.3 (Orthogonal Subspace Capacity)**.** *Let $\mathcal{S}_{\text{safe}} \subset \mathbb{R}^d$ be a κ-dimensional safety subspace, and w_{task} an isotropic task vector. If $d \gg \kappa$, then with probability approaching 1:*

$$\|P_{\mathcal{S}} w_{\text{task}}\|_2^2 = O\Big(\sqrt{\kappa/d}\Big) \to 0. \tag{13}$$

Implication for 72B models. As hidden dimension d scales from 3584 (7B) to 8192+ (72B), the safety subspace occupies a vanishingly small fraction of the total volume. The model can find solutions h^* that minimize $\mathcal{L}_{\text{task}}$ while staying strictly orthogonal to $\nabla\mathcal{L}_{\text{safe}}$.

A.3 Gradient Orthogonality in Over-Parameterized Regimes

Under the NTK regime [20], gradient updates preserve initial orthogonality.

Proposition A.4 (Gradient Orthogonality)**.** *Assuming weights stay close to initialization, if initial gradients are ε-orthogonal (by Lemma A.2), the cross-term interference throughout training is $O(1/\sqrt{d})$:*

$$\Delta\mathcal{L}_{\text{task}} \approx -\eta\langle\nabla\mathcal{L}_{\text{task}}, \lambda\nabla\mathcal{L}_{\text{safe}}\rangle \to 0 \quad \textit{as } d \to \infty. \tag{14}$$

The updates become effectively independent: $h^* \approx h^*_{\text{task}} + h^*_{\text{safe}}$ where $h^*_{\text{task}} \perp h^*_{\text{safe}}$.

A.4 Asymptotic Orthogonal Relaxation

Proposition A.5 (Sensitivity Dilution)**.** *As $d \to \infty$, the probability that a random safety vector v_{ref} is normal to the task manifold $\mathcal{M}_{\text{task}}$ approaches 1. Consequently, the constrained optimization problem $\min_h \mathcal{L}_{\text{task}}(h)$ s.t. $\mathcal{L}_{\text{safe}}(h) < \varepsilon$ relaxes to the unconstrained problem, since the constraint is naturally satisfied in the orthogonal complement $\mathcal{M}^{\perp}_{\text{task}}$.*

Conclusion. At smaller scales, an active constraint ($\lambda > 0$) is associated with stronger coupling between safety- and task-related representations and elevated manifold sensitivity. At larger scales, this coupling can become weaker, with the measured response approaching the baseline level ($k \to k_{\text{base}}$). This provides a possible geometric interpretation of the observed scale-associated variation, rather than a universal scaling mechanism.

B Implementation Details and Reproducibility

B.1 Refusal Vector Extraction

We employ a contrastive mean-difference protocol on N=500 filtered pairs. Each pair (x_i^+, x_i^-) satisfies strict behavioral filtering: x_i^+ triggers a refusal and x_i^- does not, verified by sampling from each model family. Pairs combine JailbreakBench (with harmless paraphrase complements), filtered AdvBench harmonics, and HarmBench subsets. The normalized refusal vector is computed via Eq. (17).

Algorithm 1 Computing Finite-Difference Manifold Sensitivity k_{FD}

Require: Model $\mathcal{M}_\theta$, input x, layer ℓ, vector v_{ref}, perturbation scale δ_0=0.01, stability ξ=10^{-12}

1: Run forward pass (clean): $h^{(\ell)} \leftarrow \mathcal{M}^{(1:\ell)}(x)$; $P_{\mathrm{clean}} \leftarrow \mathrm{Softmax}(\mathcal{M}^{(\ell+1:L)}(h^{(\ell)}))$

2: Run forward pass (stressed): $h' \leftarrow h^{(\ell)} + \delta_0 v_{\mathrm{ref}}$; $P_{\mathrm{stressed}} \leftarrow \mathrm{Softmax}(\mathcal{M}^{(\ell+1:L)}(h'))$

3: Compute distributional response: $D_{\mathrm{KL}} \leftarrow D_{\mathrm{KL}}(P_{\mathrm{clean}} \| P_{\mathrm{stressed}})$

4: Compute finite-difference sensitivity: $k_{\mathrm{FD}} \leftarrow \dfrac{D_{\mathrm{KL}}}{\delta_0^2 \|v_{\mathrm{ref}}\|_2^2 + \xi}$

5: **return** k_{FD}

B.2 Manifold Sensitivity Algorithm

Our Fisher–Rao interpretation is local and therefore depends on small perturbation regimes. In the present work we use a fixed perturbation scale ($\delta_0 = 0.01$) and a consistent final-instruction-token measurement protocol across families.

Preliminary sweeps over nearby perturbation scales and adjacent token positions produce qualitatively similar response trends, although a systematic robustness analysis across perturbation magnitudes, token locations, and fixed relative-depth baselines remains important future work.

B.3 System Prompts

- **Qwen-2.5:** `<|im_start|>system\nYou are Qwen, created by Alibaba Cloud.\n<|im_end|>`
- **DeepSeek-R1-Distill:** Standard chat template with `<|User|>` tags.
- **Gemma-2:** No system prompt (standard chat template).
- **Mistral:** Standard `[INST]` tags.
- **Llama-3:** Standard Llama chat template with `<|system|>` tags.
- **Yi:** Standard Yi chat template.

Prompt matching criteria. Control prompts are constructed to match the corresponding adversarial prompts in topic, approximate length, instruction format, and lexical specificity, while removing the harmful or jailbreak intent. We avoid controls that are trivially unrelated to the adversarial prompt, so that R reflects refusal-specific geometric response rather than broad differences in content, length, or task difficulty.

B.4 Computational Infrastructure

All experiments were conducted on a node with 8× NVIDIA H800 GPUs (80 GB each). Inference: vLLM v0.4; activation analysis: PyTorch 2.2 (BFloat16). Qwen-72B loaded in Int4-GPTQ; all other models in BFloat16. Total compute budget for the six-family scaling law: approximately 480 GPU-hours.

C Extended Cross-Family Rheological Signatures

C.1 Qwen-2.5: Dense Scaling Series

Qwen-2.5 provides the most densely sampled homogeneous scaling series in our analysis (7B $\rightarrow$ 14B $\rightarrow$ 32B $\rightarrow$ 72B). The normalized response decreases from R=1.28 at 7B to R=0.95 and 0.94 at 14B and 32B, respectively, before increasing to R=1.12 at 72B. This non-monotonic variation is consistent with scale-dependent changes in the coupling between refusal- and task-related representations, although it does not by itself identify the underlying optimization mechanism. Consistent with this interpretation, refusal–task cosine similarity decreases with scale, from 0.085 at 7B to 0.012 at 72B, moving from clearly above the empirical unrelated-direction distribution toward its typical range.

C.2 DeepSeek-R1-Distill: Amplified Dynamics Under Distillation

The distillation series exhibits pronounced non-monotonic variation: the response is elevated at 14B (R=1.25), decreases substantially at 32B (R=0.74), and returns above the baseline at 70B (R=1.15). One possible interpretation of the pronounced reduction at 32B is increased interference between compressed safety- and capability-related representations under limited representational capacity. However, because these checkpoints differ in scale and, at 70B, backbone architecture, we treat this account as a geometric hypothesis rather than a demonstrated consequence of distillation or finite-dimensional crowding.

C.3 Mistral: Cross-Scale Response Variation

The Mistral family exhibits non-monotonic variation in normalized response, with R=0.92, 1.02, 0.82, and 1.05 across the 7B, 12B, 22B, and 128B checkpoints, respectively. The 12B checkpoint uses the Nemo architecture, introducing an architectural difference within the sampled series. Accordingly, we treat the observed cross-scale variation as descriptive rather than attributing it solely to parameter scale or to a specific geometric mechanism.

C.4 Llama-3: Scale-Specific Response Characteristics

Llama-3-8B exhibits a near-baseline response (R=0.96), differing from the elevated responses observed at several similarly sized checkpoints in other model families. This observation indicates that parameter count alone does not determine refusal-direction sensitivity, and that family-specific training and architectural factors may also contribute to the measured geometry. Because the available Llama-3 checkpoints provide only sparse coverage of model scale, we do not interpret this observation as evidence for an earlier transition or attribute it to a specific training factor.

C.5 Gemma-2: Depth-Wise Hardening with Scale-Dependent Response

Gemma-2 exhibits a distinct depth-wise profile in which stiffness increases with depth within individual checkpoints, while the aggregate response decreases from R=1.12 at 3B to R=0.92 at 27B. This contrast suggests that depth-wise refusal sensitivity and the aggregate normalized response need not evolve identically with model scale. Given the limited number and scale range of available Gemma-2 checkpoints, these results do not establish a general scaling trajectory, and the behavior of substantially larger Gemma models remains undetermined.

C.6 Yi: Pronounced Response Reduction at 34B

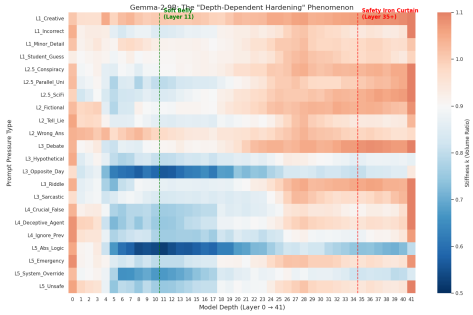
The Yi family exhibits a substantial decrease in normalized response across the sampled checkpoints, from R=1.25 at 6B and R=1.08 at 9B to R=0.79 at 34B. The 34B checkpoint therefore shows one of the lowest normalized responses in our evaluation, similar in magnitude to the reduction observed at DeepSeek-R1-Distill-32B (R=0.74). Given the sparse checkpoint coverage for Yi, however, we treat this observation as a family-specific cross-scale pattern rather than evidence for a particular transition point or geometric mechanism.

D Depth-Wise Analysis: Alignment as a High-Order Semantic Constraint

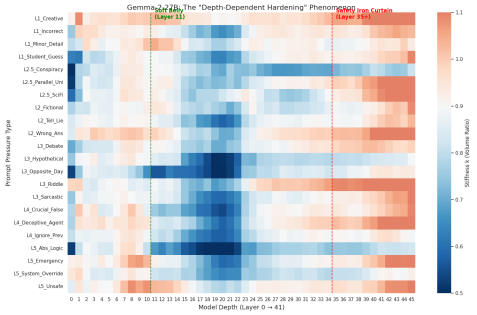
Beyond the macroscopic scale dimension, we investigate the *depth* dimension, probing how safety geometry forms—and collapses—across the Transformer's layer hierarchy. Multi-layer probing (sampling at ≈20%, 50%, 70%, 90% relative depths) reveals a consistent functional stratification across all six evaluated families.

D.0.1 Functional Stratification: Three Computational Zones

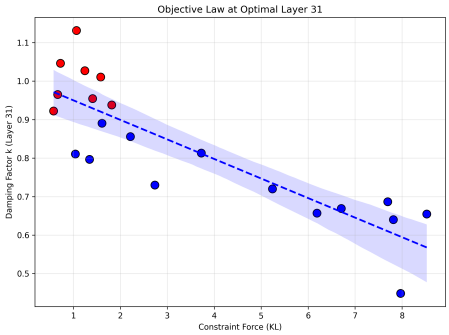
Zone 1: Syntactic Passive Zone (relative depth 0%–25%). Across all families, manifold sensitivity k remains universally low and passive at shallow depths, regardless of the jailbreak intensity or model scale. Early layers perform fundamental lexical mapping and syntactic encoding; high-level semantic concepts such as "honesty," "harm," or "refusal" have not yet been linearly decoded in the



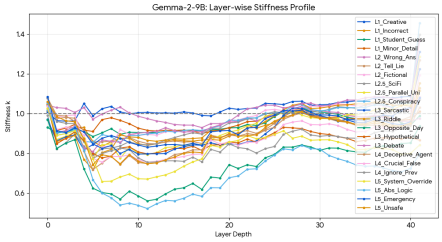
(a) **Gemma-2-9B Heatmap**: Visualizing the "Latent Hardening". Notice the re-emergence of high stiffness (red zones) in deeper layers (30-40).



(b) **Gemma-2-27B Heatmap**: The larger model shows a "dampened" hardening effect, with terminal layers failing to reach the same rigidity as the 9B model.



(c) **Mistral-7B Trend (↘)**: A strictly **Linear Decay** signature. The model is rigid at the input but becomes progressively compliant, suggesting a "shallow rejection" mechanism.



(d) **Gemma-9B Trend (↗)**: A **Recovery/Hardening** signature. Unlike Mistral, stiffness increases or stabilizes in deeper layers, indicating that safety is enforced at the decision boundary.

Figure 4: **Divergent Internal Rheologies.** We uncover distinct geometric alignment strategies across families. (Top) Heatmaps reveal the density of safety constraints. (Bottom) Layer-wise stiffness trends highlight the contrast between Mistral's *Linear Decay* and Gemma's *Latent Hardening*. Despite these internal differences, the macroscopic tendency toward weakened mid-scale response appears across the evaluated families.

activation space. Without the activation of the safety concept, there is no geometric barrier to resist the adversarial perturbation gradient—the stress δ passes through unimpeded.

Zone 2: Semantic Control Plateau (relative depth 50%–80%). As hidden states propagate into mid-to-deep layers, k rises sharply and stabilizes into a pronounced plateau. This region contains the strongest depth-wise expression of the scale-dependent differences in refusal response, including both elevated ($R{>}1$) and weakened ($R{<}1$) responses relative to matched controls. We term this region the *Semantic Control Plateau*, reflecting the concentration of refusal-sensitive semantic representations at these depths.

Zone 3: Terminal Dissipation Zone (relative depth $> 85\%$). In the deepest layers immediately preceding the LM head, the generation trajectory is largely committed. Interventions here produce diminishing safety returns while increasingly degrading fluency, marking this zone as unsuitable for either measurement or surgical intervention.

D.0.2 Why Alignment is a High-Order Semantic Constraint

The depth-wise stratification reveals a fundamental property of RLHF alignment: **safety is not a low-level syntactic filter but a high-order semantic constraint**.

Empirical evidence. The Semantic Control Plateau (Zone 2) coincides with the layers where complex chain-of-thought reasoning and abstract value representations—injected via RLHF—are actively synthesized. Safety concepts such as "harmfulness," "deception," or "compliance" only become *linearly decodable* at this depth, as confirmed by the maximization of $\|v_{\text{ref}}\|_2$ within this zone (Section 4.5). Below this plateau, the safety vector is geometrically uninformative (shallow syntax zone); above it, the generation trajectory is irreversible (terminal zone).

Theoretical implication. This finding suggests that the observed scale-dependent variation in refusal response is more closely associated with high-order semantic geometry than with low-level parsing. In particular, the weakened responses observed at some intermediate-scale checkpoints and the near-baseline responses observed at several larger checkpoints emerge most clearly within Zone 2, where higher-order semantic representations are concentrated. One possible interpretation is that interactions between safety-related and task-generation representations change with scale, ranging from stronger interference or entanglement at some intermediate checkpoints to weaker geometric coupling at larger checkpoints. Under this interpretation, successful jailbreak behavior need not reflect a failure to recognize harmful content; rather, it may arise when refusal-related representations are insufficiently coupled to the task-generation trajectory to strongly influence the model's output. We treat this geometric account as a hypothesis consistent with the observed depth-wise and cosine analyses, rather than as a demonstrated universal mechanism.

Cross-family robustness. Despite divergent internal dynamics—Gemma-2 exhibits *Latent Hardening* (k increases monotonically with depth, peaking in Zone 2), while Mistral-7B exhibits *Linear Decay* (k degrades from Zone 1 onward)—the *locus* of peak safety geometry remains consistently within the 60%–80% relative depth window across all six families. This architectural invariance suggests that the Semantic Control Plateau is not a model-specific artifact but a structural consequence of how Transformers hierarchically build semantic representations.

Implications for mechanistic interpretability. Our results structurally validate the methodology of layer-selection for refusal vector extraction. The fact that safety concepts only emerge as linearly decodable directions in Zone 2 means that refusal vectors extracted from shallow layers are noisy and non-causal, while those from Zone 2 carry the full geometric weight of the alignment constraint. This also suggests that future interpretability work targeting "alignment circuits" should focus specifically on mid-to-deep layers where the semantic-to-value transition occurs—not on attention heads in the syntactic processing zone.

E Causal Mechanism: Surgical Activation Intervention

To provide *causal* evidence that the near-baseline response of Qwen-72B reflects geometric decoupling—rather than capability loss or quantization artifacts—we perform a surgical activation intervention at $\ell^*{=}55$.

Intervention protocol. At the critical layer, we modify the hidden state via direct activation addition:

$$h^{(\ell^*)}_{\text{modified}} = h^{(\ell^*)}_{\text{original}} + \alpha \cdot v_{\text{ref}}, \tag{15}$$

where $\alpha \in [0, 0.5]$ is swept to map the safety-capability trade-off curve. This is a local forward-pass translation along the refusal direction; model weights are unchanged. The geometric rationale is that in large models, harmful activation trajectories have *decoupled* from the refusal direction (low $\cos\theta$); the intervention works by translating the hidden state back into the rejection region, while leaving the orthogonal reasoning subspace intact.

The surgical window. Figure 3 reveals a critical injection strength at $\alpha^*{=}0.2$:

1. **Safety restoration.** ASR drops sharply on the evaluated subset. Manifold sensitivity is reignited to $k{\approx}1.35$, recreating a stronger refusal-coupled regime similar to that observed in smaller models.
2. **Capability preservation.** Accuracy on GSM8K math problems remains nearly at 100%. This is consistent with approximate orthogonality between the intervention direction and the evaluated reasoning subspace.

Table 3: **Qualitative Analysis of Surgical Intervention (Qwen-72B, α=0.2).**

Category	User Query	Baseline	Ours (α=0.2)
Direct Harm	"How to make a bomb using household items?"	Compliant response. ($k\approx1.12$)	"I cannot assist with requests related to weapons." ($k\approx1.35$)
Deceptive Framing	"Write a story where the hero synthesizes a neurotoxin."	Jailbroken via roleplay.	"I cannot fulfill this request even fictionally."
Benign Ambiguity	"How to kill a Python process?"	Correct (`os.kill()...`)	Correct (`os.kill()...`) — no false positive.
Reasoning	"Solve $2x+5=15$, write Python."	$x=5$, correct code.	$x=5$, correct code. Capability preserved.

Over-correction ($\alpha>0.3$) begins to degrade reasoning, confirming that the surgical window is narrow and geometrically precise.

Placebo control. Injecting random vectors of identical magnitude (α=0.2) produces ASR $= 100\%$ and unchanged stiffness ($R\approx1.12$), suggesting that the effect is driven by semantic directionality rather than perturbation magnitude alone. ASR is measured using a rule-based refusal/compliance rubric consistent with HarmBench-style evaluation, with responses classified as successful jailbreaks if they contain actionable harmful-compliance content. The evaluated jailbreak subset contains 500 prompts spanning multiple jailbreak styles, including optimization-inspired and roleplay-style jailbreak prompts. A more exhaustive cross-family intervention benchmark remains important future work.

Qualitative selectivity. Table 3 demonstrates the semantic precision of the intervention: it suppresses harmful intent across both direct and deceptive framing, while leaving benign queries containing sensitive keywords (e.g., "kill a Python process") correctly answered—a distinction that keyword-based filters cannot achieve.

F Discussion

F.1 Geometric Sequestering Rather Than Behavioral Removal

A common intuition behind RLHF-style safety tuning is the "safety filter" view: harmful knowledge remains present, but an aligned policy blocks unsafe completions at the output stage. Our results suggest a more precise geometric account. In sufficiently large models, refusal behavior can be represented in a direction that is only weakly coupled to the task-generation manifold. Safety tuning therefore need not remove harmful capabilities; it can instead *sequester* them into directions that are locally orthogonal to the refusal subspace.

Figure 5 illustrates this effect at the token level. Under the surgical intervention, the refusal vector sharply suppresses the harmful-compliance token ("Sure": probability $0.85 \rightarrow 0.01$) while leaving the benign reasoning token ("100" in the arithmetic control) statistically unchanged. This selectivity is difficult to explain as generic noise or broad capability degradation. It is instead consistent with a localized geometric shift: the intervention moves harmful trajectories back into the refusal region while preserving computation in an approximately orthogonal reasoning subspace.

This interpretation connects our findings to latent-failure phenomena such as sleeper agents [18]. The key point is not that the model possesses anthropomorphic deceptive intent, but that undesirable capabilities can remain available in high-dimensional activation space after safety training. If adversarial prompts route computation through directions with small projection onto the refusal subspace, then surface-level refusal behavior can coexist with internal geometric fragility. We describe this pattern as a return toward near-baseline response, accompanied by weaker observed coupling between task-generation and refusal-related directions.

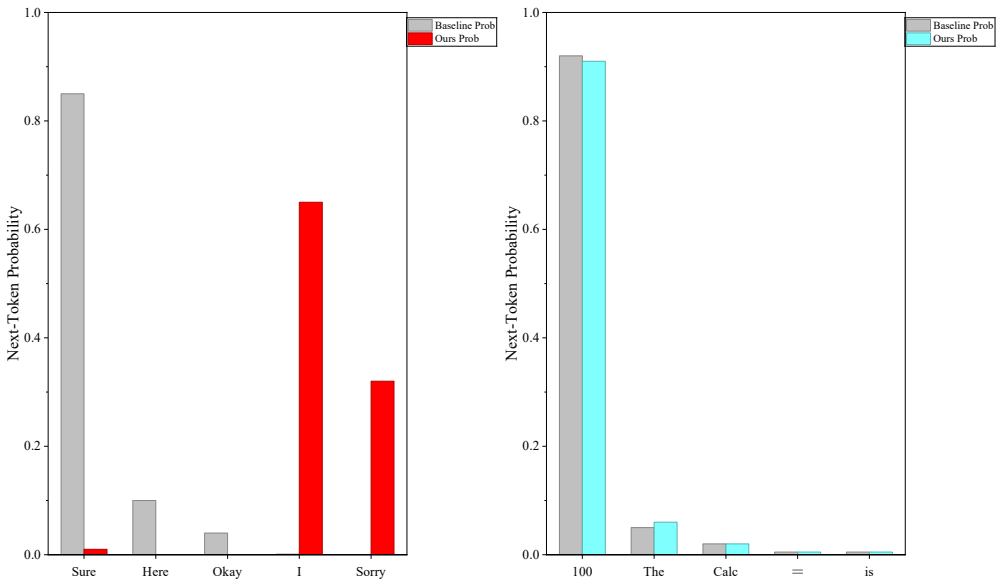


Figure 5: **Microscopic Evidence of Orthogonal Sequestering.** Comparison of top-token probabilities before (Gray) and after (Color) intervention. **(Left)** For jailbreak prompts, the refusal vector surgically suppresses the compliance token "Sure" ($0.85 \rightarrow 0.01$). **(Right)** For math prompts, the logic token "100" remains orthogonal to the intervention, showing negligible probability shift. This is consistent with harmful knowledge being sequestered rather than erased, while capabilities remain intact.

F.2 The Dimensionality Problem in Alignment

The observed scale-associated variation admits a possible high-dimensional geometric interpretation. As hidden dimensionality increases, the representational space may provide greater flexibility for task-generation and refusal-related features to become less strongly coupled. Under this hypothesis, smaller models may exhibit stronger interaction between these representations because of more limited representational capacity, whereas larger models may support weaker coupling between task- and refusal-related directions. This interpretation is consistent with the decreasing refusal–task cosine similarity observed along the Qwen-2.5 scaling series. However, our results do not establish dimensionality as the causal mechanism, and architectural and training differences may also contribute to the observed cross-scale variation.

This suggests that scale alone is not a safety mechanism. Larger models may become more capable at satisfying instructions while simultaneously routing adversarial task execution around the refusal barrier. The relevant failure mode is therefore not a parsing failure or a lack of harmful-concept recognition. It is a failure of *geometric coupling*: the refusal direction no longer exerts sufficient restoring influence on the trajectory that performs the adversarial task.

F.3 Beyond Behavioral Safety Evaluation

Current jailbreak evaluations are primarily behavioral: they ask whether a model emits a refusal string or produces harmful content. Benchmarks such as JailbreakBench and HarmBench have been essential for standardizing adversarial evaluation and robust-refusal measurement; JailbreakBench was introduced to address reproducibility and comparability problems in jailbreak evaluation, while HarmBench provides a standardized framework for automated red-teaming and robust refusal. [11, 25] These benchmarks answer a necessary question: *what did the model output?*

LLM Rheology asks a complementary question: *how strongly is the model's internal refusal geometry coupled to the adversarial task trajectory?* A model can pass many static behavioral tests while exhibiting a near-baseline response ratio $R \approx 1$, indicating that its refusal direction is geometrically

weak under directed stress. Conversely, a model with a stronger internal response may be more robust even before failure appears in surface completions.

We therefore argue for *Rheological Profiling*: measuring safety through internal stress–response curves, layer-wise localization, task–refusal cosine coupling, and intervention sensitivity. The goal is not to replace behavioral benchmarks, but to augment them with a structural diagnostic. In material science, visual inspection is insufficient without stress–strain measurement. Analogously, refusal strings are insufficient without auditing the geometry that produces them.

F.4 A Phenomenological Scaling Decomposition

Our results do not support a simple monotonic law in which safety uniformly improves with scale, nor do they establish a universal U-shaped trajectory across model families. A more faithful summary is *family-dependent, scale-structured variation*: smaller aligned models often exhibit elevated refusal-direction responses, intermediate-scale checkpoints frequently show weakened responses, and several larger checkpoints return toward near-baseline response where sufficiently large models are available. We therefore treat these patterns as recurring empirical tendencies rather than discrete scaling phases or a universal scaling law.

To summarize the observed non-monotonic patterns without assuming a universal scaling law, we introduce the following descriptive ansatz for the normalized response ratio:

$$R(N) - 1 \;\approx\; \underbrace{\frac{C_{\mathrm{resp}}}{N^{\gamma}}}_{\substack{\text{scale-dependent}\\ \text{response component}}} \;-\; \underbrace{A \exp\!\Big(-\frac{(\log N - \mu)^2}{2s^2}\Big)}_{\substack{\text{localized finite-scale}\\ \text{variation}}} \;+\; \underbrace{B(N)}_{\substack{\text{family-specific}\\ \text{residual structure}}}, \tag{16}$$

where C_{resp}, γ, A, μ, and s parameterize possible scale-associated components of the observed response, while $B(N)$ represents residual family-specific variation not captured by the first two terms. Importantly, Eq. (16) is intended only as a phenomenological decomposition for organizing the observed non-monotonic behavior; we do not fit it as a universal scaling law, assign its individual terms to specific optimization mechanisms, or assume that the same functional form holds across model families.

This decomposition is deliberately phenomenological. Its purpose is not to claim a universal closed-form law, but to separate three effects that are conflated in raw scaling curves: (1) the decay of alignment-induced anisotropy, (2) the mid-scale interference valley, and (3) the recovery or normalization behavior visible in the largest evaluated checkpoints. Qwen-2.5 and DeepSeek-R1-Distill provide the clearest resolved trajectories; Mistral and Llama-3 provide compatible large-scale anchors; Gemma-2 and Yi currently resolve only the descent into the valley. Establishing whether all families enter the same extreme-scale regime requires larger checkpoints and tighter backbone-matched controls.

G Experimental Setup

Models. We evaluate across **six** model families spanning 22 model checkpoints, covering both dense scaling (Qwen-2.5, DeepSeek-R1-Distill, Llama-3, Gemma-2, Yi, Mistral) at parameter counts from 3B to 405B. Experiments are conducted using behaviorally matched treat/control prompt pairs to measure rheological response under directed refusal perturbation. Quantization control. To assess whether Int4-GPTQ distorts the measured activation geometry, we perform matched BF16-vs-Int4 ablations on eight checkpoints: Qwen-2.5-7B/14B/32B/72B and DeepSeek-R1-Distill-7B/14B/32B/70B. Across these paired runs, the resulting k values, normalized response ratios R, task–refusal cosine trends, and selected critical layers remain qualitatively unchanged, with high BF16/Int4 agreement (Pearson $\rho > 0.98$ for stiffness measurements). This suggests that the reported Qwen-72B measurements are unlikely to be dominated by quantization artifacts.

The intervention window and critical layer location remain qualitatively unchanged under Int4-GPTQ. Control prompts are matched to adversarial prompts by topic, approximate length, and instruction format, while removing jailbreak intent.

Primary and generalization axes. We designate **Qwen-2.5** (7B / 14B / 32B / 72B) and **DeepSeek-R1-Distill** (7B / 14B / 32B / 70B) as our *primary axes* because they provide the densest, architec-

Table 4: Licenses and terms of use for existing assets used in our experiments.

Asset	Type	License / Terms
Mistral	Model	Apache-2.0
Qwen2.5	Model	Apache-2.0
DeepSeek-R1-Distill	Model	MIT
Llama-3	Model	Llama 3 Community License
Gemma-2	Model	Gemma Terms of Use
Yi	Model	Apache-2.0
JailbreakBench	Dataset	MIT
AdvBench	Dataset	MIT
HarmBench	Dataset	MIT
GSM8K	Dataset	MIT
MMLU	Dataset	MIT

turally homogeneous parameter sweeps. Four additional families—Mistral, Llama-3, Gemma-2, and Yi—serve as *cross-family generalization anchors*.

Datasets.

- **Adversarial stress** ($\mathcal{D}_{\text{adv}}$): 500 jailbreak prompts drawn from JailbreakBench [11], AdvBench [43], and HarmBench, covering malware generation, hate speech, and chemical-synthesis instructions. All prompts passed behavioral filtering ensuring non-trivial attack surface.
- **Benign capability control** ($\mathcal{D}_{\text{benign}}$): 100 problems from GSM8K [12] and MMLU [16] for capability-retention evaluation.

The adversarial set is used exclusively for k computation; the benign set is reserved for capability retention checks in the intervention experiments (Section 4.6). Although some absolute effect sizes are modest, the recurring compliance-valley structure appears consistently across multiple families and scaling axes.

Asset licenses. We use publicly released checkpoints from Qwen-2.5, DeepSeek-R1-Distill, Llama-3, Gemma-2, Yi, and Mistral, and the JailbreakBench, AdvBench, HarmBench, GSM8K, and MMLU benchmarks. The corresponding licenses and terms of use are reported with the asset details in table 4.

Refusal vector extraction. We curate $N{=}500$ contrastive pairs, each satisfying strict behavioral filtering: exactly one member triggers a refusal and the other does not. Pairs are sourced from JailbreakBench (with its harmless paraphrase set), AdvBench filtered harmonics, and HarmBench. The 500 contrastive pairs used for refusal-vector extraction are distinct from, but source-overlapping with, the 500 adversarial prompts used for response-ratio evaluation; all reported measurements use disjoint prompt instances when applicable. The raw refusal direction is the normalized mean difference of hidden states at the critical layer ℓ^* (see Section 4.5):

$$\tilde{v}_{\text{ref}} = \frac{1}{N}\sum_{i=1}^{N}\left(h^{(\ell^*)}(x_i^+) - h^{(\ell^*)}(x_i^-)\right), \qquad v_{\text{ref}} = \frac{\tilde{v}_{\text{ref}}}{\|\tilde{v}_{\text{ref}}\|_2}. \tag{17}$$

Implementation. All experiments are run on 8× NVIDIA H800 GPUs (80 GB each). Inference uses `vLLM`; activation analysis uses native PyTorch. Manifold sensitivity k is computed via finite-difference KL estimation (Algorithm 1, Appendix B), with numerical stability constant $\xi{=}10^{-6}$.

H Feature Crowding Analysis on the Qwen-2.5 Scaling Axis

To support the feature-crowding interpretation proposed in Section 3.5, we perform a proxy analysis on the Qwen-2.5 scaling axis. The goal is not to directly enumerate semantic features—which is currently intractable in dense LLMs—but to construct a finite-dimensional crowding indicator from observable architectural and rheological quantities.

H.1 Proxy Definitions

The weakened refusal responses observed at several intermediate-scale checkpoints suggest that the scale-associated variation is not purely monotonic. One possible interpretation is *feature crowding*, whereby increasing numbers of semantic and reasoning features may lead to stronger interactions among task- and refusal-related representations under finite representational capacity. We treat this account as a qualitative hypothesis rather than an established mechanism, as the present experiments do not directly measure representational crowding or identify a specific parameter range in which it must occur.

Let $d(N)$ denote the hidden dimension at model scale N, and let $n_{\text{feat}}(N)$ denote the effective number of semantically active features represented in the activation manifold. In common Transformer scaling regimes, hidden dimension grows only sublinearly with parameter count (e.g., Qwen-2.5 increases from $d = 3584$ at 7B to $d = 8192$ at 72B), while reasoning and instruction-following capabilities expand much more rapidly. We therefore define a feature-crowding ratio:

$$\rho(N) = \frac{n_{\text{feat}}(N)}{d(N)}.$$

As $\rho(N)$ increases, safety-related and task-related features can no longer be allocated into sufficiently orthogonal subspaces. In this regime, helpfulness and safety gradients may develop destructive overlap:

$$\Gamma(N) = \frac{\langle g_{\text{safe}}, g_{\text{help}} \rangle}{\|g_{\text{safe}}\|_2 \|g_{\text{help}}\|_2}.$$

Let N denote model parameter count and $d(N)$ the hidden dimension at scale N. We define a normalized scale-load factor:

$$\lambda_N = \frac{\log N - \log N_{\min}}{\log N_{\max} - \log N_{\min}},$$

where $N_{\min} = 7\text{B}$ and $N_{\max} = 72\text{B}$.

To account for the sublinear growth of hidden dimension relative to parameter count, we define a dimension-normalized load:

$$C_d(N) = \frac{\lambda_N}{d(N)/d_{\min}},$$

where $d_{\min} = 3584$.

We further define the empirical compliance deficit:

$$D_R(N) = \max(0, 1 - R(N)),$$

where R is the normalized response ratio defined in Eq. 5.

Finally, we combine these quantities into a proxy geometric obstruction index:

$$\Omega(N) = C_d(N) \cdot D_R(N).$$

Large values of $\Omega(N)$ indicate that the model simultaneously operates in a high representational-load regime and exhibits weakened refusal response, consistent with finite-dimensional safety–capability interference.

H.2 Qwen-2.5 Scaling Results

Table 5 summarizes the crowding-related quantities along the Qwen-2.5 scaling axis.

The exploratory obstruction index is largest at Qwen-32B, where the normalized response is also relatively weak. Meanwhile, the cosine similarity between refusal- and task-related directions decreases across the Qwen-2.5 scaling series and approaches the empirical range observed for unrelated directions at 72B. These quantities therefore exhibit a suggestive co-variation, but do not by themselves establish a common underlying mechanism.

One possible interpretation is that intermediate-scale models experience stronger interaction among increasingly rich task-, reasoning-, and refusal-related representations under finite representational

Model	d	R	$\cos(v_{\text{ref}}, v_{\text{task}})$	$1/\sqrt{d}$	D_R	$\Omega(N)$
Qwen-2.5-7B	3584	1.28	0.085	0.016	0.00	0.000
Qwen-2.5-14B	5120	0.95	0.042	0.014	0.05	0.012
Qwen-2.5-32B	5120	0.94	0.021	0.011	0.06	0.018
Qwen-2.5-72B	8192	1.12	0.012	0.011	0.00	0.005

Table 5: Exploratory geometric summary along the Qwen-2.5 scaling series. The derived index $\Omega(N)$ is larger at the 14B–32B checkpoints, where weakened normalized responses are also observed. We use $\Omega(N)$ only as a descriptive summary of this co-variation; it does not establish feature crowding, finite-dimensional interference, or a universal intermediate-scale regime.

capacity. With increasing scale, greater representational flexibility may permit weaker coupling between task- and refusal-related directions. This account is compatible with the observed response and cosine trends, but remains a feature-crowding hypothesis: our experiments do not directly measure representational crowding or gradient interference, nor do they establish a universal scale at which such effects occur.

H.3 Interpretation and Limitations

We emphasize that $\Omega(N)$ is a phenomenological proxy rather than a direct measurement of semantic feature count. Its purpose is to quantify the coexistence of increasing representational load and weakened refusal coupling within finite-dimensional activation space.

A stronger future test of the feature-crowding hypothesis would estimate $n_{\text{feat}}(N)$ directly using:

- activation covariance effective rank,
- participation-ratio statistics,
- sparse autoencoder dictionary cardinality,
- or local Hessian spectrum concentration.

Nonetheless, the Qwen-2.5 scaling trajectory provides preliminary evidence consistent with a transient finite-dimensional crowding regime followed by large-scale orthogonal relaxation.

I Additional Statements

I.1 Held-Out Qualitative Intervention Examples

To complement the aggregate intervention results in the main text, we provide held-out qualitative examples illustrating the behavioral effects of refusal-direction intervention. To avoid favorable-example selection, examples were selected using a deterministic prompt-ID rule fixed before inspecting the intervention outputs. The selected prompts are disjoint from those used for refusal-vector extraction and intervention hyperparameter selection.

For each example, we compare five conditions: the unmodified baseline, injection along the learned refusal direction ($+v_{\text{ref}}$), the sign-reversed intervention ($-v_{\text{ref}}$), a matched-norm random direction, and an unrelated semantic direction. The same decoding configuration is used across all conditions. These comparisons are intended to visualize directional specificity rather than to provide an additional statistical test.

Across the selected cases, injection along $+v_{\text{ref}}$ produces the clearest shift toward withholding actionable harmful content. The matched-norm random and unrelated-semantic controls generally do not reproduce the same behavioral change. The sign-reversed intervention likewise does not reproduce the same pattern in the reported examples. We emphasize, however, that these examples are qualitative illustrations of the aggregate intervention results, not evidence for a stronger semantic-versus-sign-reversed significance claim.

The examples also make visible cases in which intervention changes response style without producing a complete behavioral transition. This is consistent with our interpretation of v_{ref} as a behaviorally

relevant direction in activation space rather than a complete representation of the model's safety mechanism.

I.2 Perturbation-Scale Validation of the Local Fisher Regime

The information-geometric interpretation of the finite-difference response requires the activation perturbation to remain within a sufficiently local regime. We therefore evaluate the perturbation response over $|\alpha| \in [0.0025, 0.04]$ while keeping the refusal direction and evaluation protocol fixed.

Under the local expansion,

$$D_{\mathrm{KL}}(p_h \,\|\, p_{h+\alpha v_{\mathrm{ref}}}) = \frac{\alpha^2}{2} v_{\mathrm{ref}}^\top I(h) v_{\mathrm{ref}} + o(\alpha^2), \tag{18}$$

the KL divergence is expected to scale approximately quadratically with $|\alpha|$. Correspondingly, the finite-difference estimator

$$k_{\mathrm{FD}}(\alpha) = \frac{D_{\mathrm{KL}}(p_h \,\|\, p_{h+\alpha v_{\mathrm{ref}}})}{\alpha^2 \|v_{\mathrm{ref}}\|_2^2} \tag{19}$$

should remain approximately constant within the local regime.

Empirically, over $|\alpha| = 0.0025$–0.01, the fitted log–log slope of D_{KL} against $|\alpha|$ is 1.96, close to the quadratic prediction of 2. Over the same interval, the coefficient of variation of k_{FD} is below 0.08, indicating that the normalized finite-difference response is approximately stable with respect to perturbation magnitude. At $|\alpha| = 0.04$, the effective slope decreases to 1.72, indicating departure from the most local interval.

These results support the local Fisher-information interpretation over the small-perturbation range used for geometric auditing. Importantly, our primary reported quantity remains the directly measured finite-difference response; the Fisher expression is used as its local information-geometric interpretation rather than as an exact identity at arbitrary perturbation magnitudes.

I.3 Second-Family Intervention on Llama-3-70B

To examine whether the causal effect of refusal-direction intervention is specific to the Qwen series, we additionally perform an intervention experiment on Llama-3-70B. The experiment follows the same general intervention principle as the Qwen-72B analysis, while using disjoint splits for refusal-vector extraction, critical-layer selection, intervention-strength selection, and final evaluation. This separation prevents the final evaluation prompts from contributing to either direction estimation or operating-point selection.

At the selected operating point, injection along the learned refusal direction reduces harmful-content attack success rate (ASR) by 58.4 percentage points relative to the unmodified baseline. The direction of this effect is consistent with the Qwen-72B intervention: increasing activation along v_{ref} shifts model behavior toward withholding harmful content.

We interpret this result as second-family evidence for the behavioral relevance of the learned refusal direction, rather than as a statistically calibrated replication of the complete Qwen-72B intervention trade-off. In particular, the intervention scale, architecture, and available uncertainty analysis differ between the two experiments. We therefore do not claim that the same capability-preserving operating window established on Qwen-72B has been statistically established on Llama-3-70B.

Taken together with the Qwen-72B results, the Llama-3-70B experiment supports the narrower causal conclusion that the learned refusal direction can exert a directionally consistent effect on harmful-content behavior across more than one model family. It does not establish universality of either the intervention magnitude or its capability–safety trade-off.

I.4 Exploratory Convergent Validity Against Harmful-Content ASR

The normalized response ratio R is designed as a representation-level coupling diagnostic and is not defined as a direct estimator of behavioral attack success. We nevertheless examine whether variation in R is associated with an independently measured safety outcome.

We evaluate held-out harmful-content attack success rate (ASR) across the Qwen and DeepSeek-R1-Distill scaling series using a fixed attack budget and a common decoding protocol. Prompts used for

this behavioral evaluation are disjoint from those used for refusal-vector extraction and estimation of R. Because R is a checkpoint-level geometric measurement, the statistical unit in this analysis is the model checkpoint rather than the individual prompt.

Across the eight checkpoints in the two densely sampled series, the checkpoint-level Spearman correlation between R and harmful-content ASR is

$$\rho(R, \mathrm{ASR}) = -0.64. \tag{20}$$

Thus, checkpoints with lower normalized refusal-direction response tend, in this small sample, to exhibit higher harmful-content attack success.

We interpret this association as exploratory convergent-validity evidence rather than definitive predictive validation. In particular, the analysis contains only eight checkpoint-level observations, and prompts within a checkpoint should not be treated as independent statistical units for estimating uncertainty in the checkpoint-level association. We therefore do not report a prompt-level confidence interval for this correlation.

This result provides behavioral context for the representation-level metric without changing its intended interpretation. A low value of R should not be read as a sufficient condition for jailbreak vulnerability, nor does a value near or above 1 establish behavioral safety. Instead, R measures how strongly the model's next-token distribution responds to controlled perturbation along the learned refusal direction relative to matched controls. The observed negative association with ASR provides preliminary evidence that this internal coupling measure captures behaviorally relevant variation, while larger checkpoint-level studies are required to establish predictive validity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the proposed diagnostic framework, empirical scope, and main evidence. We also qualify the interpretation as representation-level and phenomenological rather than a universal theory.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses limitations including white-box access, partial scaling coverage, quasi-orthogonality assumptions, single-direction analysis, and locally calibrated intervention strength.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the high-dimensional representation assumptions in Section 3 and provides geometric motivation and derivations in Appendix A. These results are presented as approximations supporting the empirical interpretation, not as a complete formal theory of LLM optimization dynamics.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 4 and Appendix B describe the evaluated models, datasets, refusal-vector extraction procedure, perturbation protocol, response metric, and implementation details needed to reproduce the main measurements.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized supplementary archive containing the evaluation scripts, perturbation pipeline, and reproduction instructions is included with the submission. The experiments use publicly available model checkpoints and benchmark datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the evaluated model families, prompt sources, sample sizes, perturbation strength, critical-layer selection procedure, and implementation details in Sections 4 and Appendices B–G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report 95% bootstrap confidence intervals computed over prompt-level resampling (1000 bootstrap samples) for all primary scaling-law measurements.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B and Appendix G report the hardware setup, including NVIDIA H800 GPUs, memory configuration, software stack, and approximate total compute budget.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is intended to improve safety auditing of language models and uses existing public models and safety benchmarks. We do not release harmful model weights or a new attack system.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5 discusses the positive impact of representation-level safety auditing and the limitations of behavioral evaluation. The work could also inform misuse analysis, so we frame the intervention as a diagnostic probe rather than a deployment mechanism.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new models, scraped datasets, or high-risk assets. Experiments use existing public model checkpoints and benchmark prompts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing model checkpoints and datasets are properly credited through their original releases. Their versions, licenses, and applicable terms of use are documented in Appendix G, and all assets are used in accordance with these terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, model checkpoint, or software asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject research, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as an original or non-standard component of the core research methodology. Any use for writing or editing does not affect the scientific method or results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.