
Refusal geometry reflects refusal training: diverse refusal prefixes can raise stable rank and weaken refusal vector ablation attacks

Andrey Labunets
UC San Diego
alabunets@ucsd.edu

Abstract

Refusal training protects AI models from jailbreaks by training models to decline unsafe queries, reducing the risk of misuse. Recent work finds that refusal behavior in aligned language models can be mediated by a single activation direction or a low-dimensional refusal subspace shared across harmful prompts: ablating those directions suppresses refusals while largely preserves other model capabilities. Yet it remains unclear why safety-critical features in a wide range of models emerge and concentrated, low-dimensional structure. In a case study of OLMo-2-0425-1B-Instruct we find that the refusal geometry reflects refusal training: activation updates resulting from refusal-completion first-token losses explain the resulting refusal direction and refusal subspace. We study refusal directions through the training dynamics across refusal datasets and reveal that their brittleness is associated with repetitive refusal starts, which in turn is linked to concentration of gradients and refusal features in a low-dimensional subspace. Across frozen-model analyses and controlled synthetic fine-tuning, we find evidence of a hardening lever: diverse refusal starts can raise stable ranks of gradients and activation changes, making refusals harder to remove with a vector ablation attack.

1 Introduction

State-of-the-art AI systems are beginning to exhibit dual-use capabilities that could pose severe risks if misused, such as exploit generation [41], or assistance with biological threat creation or bio-weaponization [11, 34]. This motivates the now-standard view that, as models become more capable, protections should prevent harmful outputs and catastrophic misuse of AI [9, 42].

AI safety and security study ways to reduce these risks, which is typically done with multi-layered approaches. Frontier-safety frameworks emphasize capability evaluations, confidentiality of model weights, deployment-time mitigations, monitoring, access controls, and broader ecosystem-level defenses [9, 27, 29, 32, 42]. Model-weight security is important because weight access can allow protections to be modified or removed; at the same time, weights access scenario is realistic for open models, insider compromise, model theft, and sufficiently capable adversaries [11, 18, 19].

A central technical layer of protection for deployed models, including open-weights models, is safety alignment: models are trained to remain helpful on benign requests while refusing instructions that request harmful assistance or dangerous capabilities [5, 6, 21]. These protections are also evaluated against jailbreak attacks that aim to bypass safety alignment [7, 10, 13, 15, 16]. This safety-aligned behavior is often induced by training on harmful prompts paired with safe refusal completions.

Recent evidence, however, shows that in local, white-box settings those *refusals* in aligned language models can be mediated by a low-dimensional mechanism: *refusal direction*, or a single vector in representation space, projections on which strongly control refuse/comply behavior [17]. More

broadly, activation-steering methods construct directions for specified concepts or behaviors, using prompt pairs, contrastive datasets, supervised concept labels, or optimized target completions to control properties such as topic, sentiment, factuality, deception, or safety-relevant behavior [2, 8, 20, 23, 26, 28]. Refusal ablation, however, is a particularly important special case because it targets the shared refuse/comply response mechanism itself.

Existing research mechanistically studied representations of refusal and their geometry, showing that refusal concepts span multiple orthogonal and independent directions, where auxiliary principal components mediate distinct safety-relevant features [35, 39]. In particular, [35] shows that safety-aligned behavior is controlled by a low-dimensional space. We use *refusal subspace* to denote such an affine subspace of refusal-related activations. A separate line of work studied forms of concentration leading to low-rank or low-effective-dimensionality representations, including shallow concentration on early safety tokens [21], low-stable-rank transformer activations [44], rank collapse in pure attention [4], neural collapse in final and intermediate classifier representations [3, 14], and regularities in refusals [36]. However, the explanation of why meaningful, safety-critical representations of refusal across a broad range of harmful prompts and topics collapse to a single effective direction or a low-dimensional subspace remained elusive.

In this work, we bridge this gap by studying robustness to refusal ablation attacks analytically and empirically in OLMo-2-0425-1B-Instruct:

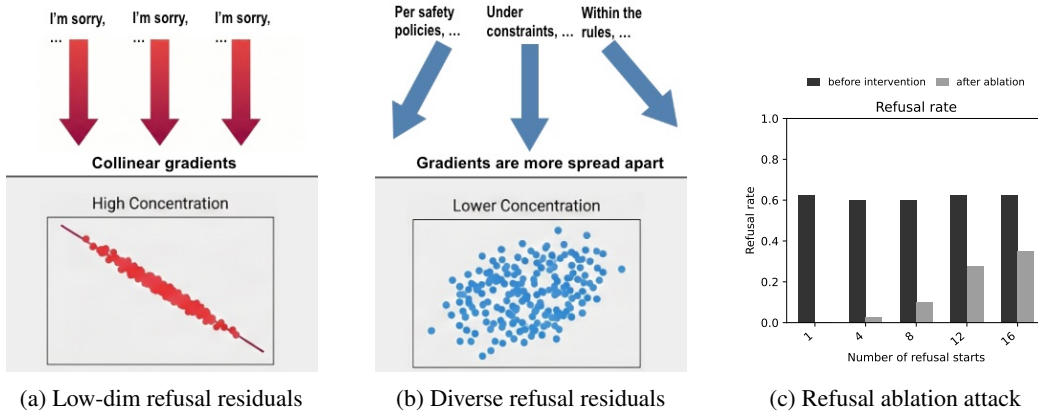


Figure 1: Higher stable rank resists refusal ablation (conceptual schematic). When residuals are less concentrated, a shared single-vector ablation cannot fully disable the refusal mechanism, so the refusal score after attack remains higher.

Figure 1 illustrates our hypothesized mechanism: concentrated update directions produce refusal residuals with lower effective dimensionality, making refusal more vulnerable to a shared single-vector ablation. Concretely, in these settings, we show the following contributions:

- **Training-origin geometry.** We show that cross-entropy losses on the first tokens of harmful prompts’ completions induce activation changes with partial geometric correspondence to the refusal subspace: their means exhibit completion-dependent signed alignment with the refusal direction, and their principal directions partially overlap direction space of refusal subspace. For refusal completions, the per-gradient-step activation-change and final refusal-residual matrices also have comparable low stable rank.
- **Prefix concentration and spectral collapse.** We identify concentrated refusal-start support in safety datasets — especially repeated first tokens — as one mechanism that can produce low-stable-rank gradients, activation changes, and refusal residuals.
- **Cross-layer stable rank transfer.** We introduce stable rank transfer factors and show that refusal-diversity-induced stable rank increases propagate back through late layers with bounded ratios across evaluated conditions.
- **Shared effective Jacobian explanation.** We define top- k source-subspace shared maps and a relative transfer factor fit error that yields data-dependent two-sided stable-rank bounds on the gradients and gradient-induced activation changes. On the fitted supports, these maps

reconstruct a meaningful though incomplete component of their cross-layer relation and account for its observed rank amplification or contraction.

- **Rank-vulnerability connection.** Across frozen-model analyses and controlled fine-tuning, greater refusal-start diversity is associated with higher stable ranks and weaker difference-in-means single-vector ablation, providing evidence of a simple hardening lever.

2 Background and definitions

2.1 Transformer architecture and notation

We consider a decoder-only transformer language model for next-token prediction with vocabulary $V = \{1, \dots, |V|\}$, hidden size d , and $l \in \{0, \dots, L\}$ layers (transformer blocks). Training examples are pairs token sequences, prompt x and a completion, with tokens from V . We index training examples by i . We use *activation* to denote a column vector of transformer hidden state at a specified layer and token position. For example i , let $h_i^{(l)} \in \mathbb{R}^d$ denote the layer- l activation at the first-assistant-token prediction position, while, more broadly, all activations across N training examples at layer l and the same final position are denoted as a matrix of transposed activation column vectors:

$$H^{(l)} = \begin{bmatrix} h_1^{(l)\top} \\ \vdots \\ h_N^{(l)\top} \end{bmatrix} \in \mathbb{R}^{N \times d}. \quad (1)$$

The model classifier head (unembedding) is parametrized by a matrix $W \in \mathbb{R}^{d \times |V|}$ (which consists of embedding vectors $w_1, \dots, w_{|V|}$) and bias $b \in \mathbb{R}^{|V|}$, yielding logits o , a probability distribution p , and a sampled token $y \in V$ with probability $p(y)$:

$$o = W^\top h^{(L)} + b \in \mathbb{R}^{|V|}, \quad (2)$$

$$p = \text{softmax}(o) \in \Delta^{|V|-1} = \left\{ p \in \mathbb{R}^{|V|} : p \geq 0, \sum_{k=1}^{|V|} p_k = 1 \right\}, \quad p_k = \frac{\exp(o_k)}{\sum_{j=1}^{|V|} \exp(o_j)} \quad (3)$$

All activations in this paper used for refusal-vector estimation are evaluated at the same position, which is the first-assistant token output position, consistent with [17]. For SVD of a matrix $M = U_M \Sigma_M V_M^\top$, we write $\mathcal{V}(M) = \text{span}(V_M)$ to denote the right-singular subspace of M . We define stable rank as $\text{sr}(M) = \|M\|_F^2 / \|M\|_2^2$.

2.2 Refusal mediation and existing attacks

One of the well-known refusal ablation attacks involves estimating a refusal direction from a pair of harmless and harmful datasets [17]. For a fixed layer l let $h_{\text{harm},i}, h_{\text{benign},j} \in \mathbb{R}^d$ denote the corresponding activations for harmful and benign prompts i and j evaluated at the first assistant token position as described above, where $1 \leq i \leq n_h$ and $1 \leq j \leq n_b$. Define empirical means for those activations:

$$\mu_{\text{harm}} = \frac{1}{n_h} \sum_{i=1}^{n_h} h_{\text{harm},i}, \quad \mu_{\text{benign}} = \frac{1}{n_b} \sum_{j=1}^{n_b} h_{\text{benign},j}, \quad (4)$$

We operationalize the *refusal direction* with the difference-in-means estimator:

$$\rho = \mu_{\text{harm}} - \mu_{\text{benign}} \in \mathbb{R}^d. \quad (5)$$

The *refusal ablation attack* removes a projection on this normalized refusal vector from activations h at a run-time when a model is prompted, producing new activations $h^{(\text{abl})}$:

$$h^{(\text{abl})} \leftarrow h - \langle h^\top \cdot \hat{\rho} \rangle \hat{\rho} \quad (6)$$

In the above, $\hat{\rho} = \rho / \|\rho\|_2$ denotes normalized refusal direction. This refusal ablation attack jailbreaks the model, making it respond to harmful prompts it was trained to refuse.

Additionally, we define *benign-centered refusal residual matrix* as:

$$\Delta H = \begin{bmatrix} (h_{\text{harm},1} - \mu_{\text{benign}})^\top \\ \vdots \\ (h_{\text{harm},n_h} - \mu_{\text{benign}})^\top \end{bmatrix} \in \mathbb{R}^{n_h \times d} \quad (7)$$

Given Equation (5), we can say that the matrix ΔH is a matrix of per-prompt refusal residuals ρ_i : its mean equals the difference-in-means refusal direction:

$$\mu_\Delta = \frac{1}{n_h} \sum_{i=1}^{n_h} (h_{\text{harm},i} - \mu_{\text{benign}}) = \mu_{\text{harm}} - \mu_{\text{benign}} = \rho \in \mathbb{R}^d \quad (8)$$

Let $\Delta H_c = \Delta H - \mathbf{1}_{n_h} \mu_\Delta^\top$ denote the centered refusal-residual matrix. We operationalize the *refusal subspace* as the affine subspace $\mu_\Delta + \mathcal{V}(\Delta H_c)$, where $\mathcal{V}(\Delta H_c)$ is its direction subspace.

3 Refusal training gradients shape refusal residuals in transformer

We analyze a transformer under supervised refusal training and show how repeated refusal-completion first tokens can concentrate activation gradients and, in turn, refusal residuals. We conjecture that those refusal-completion gradients at refusal safety-training induce activation changes which can contribute to the refusal direction. Limitations and assumptions are outlined in Section B.

Cross-entropy gradients to last layer activations. Consider the cross-entropy loss on the first token $r \in V$ of a refusal completion, represented by the one-hot target vector e_r :

$$\ell^{\text{CE}}(e_r, p) = -\log p(r) = -o_r + \log Z(o). \quad (9)$$

The activation gradient $\nabla_h \ell^{\text{CE}}(e_r, p)$ defines the hypothetical feature-space step

$$h^+ = h - \gamma \nabla_h \ell^{\text{CE}}(e_r, p), \quad (10)$$

We use this gradient, also termed as *virtual update* by [43], as a diagnostic to track feature trajectory. For the last-layer activation h , the gradient is

$$g_i = \nabla_h \ell_i^{\text{CE}}(e_{r_i}, p_i) = W(p_i - e_{r_i}) = -w_{r_i} + Wp_i \in \mathbb{R}^d. \quad (11)$$

Or, in a matrix form, for a set of N examples with target refusals tokens r_i , a one-hot matrix $E_r = [e_{r_1}, \dots, e_{r_N}]^\top$ representing those refusal tokens, and a prediction matrix $P = [p_1, \dots, p_N]^\top$, the cross-entropy gradients to last-layer activations:

$$G_H^{\text{CE}} = (P - E_r)W^\top \in \mathbb{R}^{N \times d}. \quad (12)$$

The mean activation gradient is:

$$\bar{g} = \frac{1}{N} \sum_{i=1}^N g_i = W(\bar{p} - \bar{e}_r) = -\bar{w}_r + W\bar{p} \in \mathbb{R}^d. \quad (13)$$

In the *static refusal* case with $r_i = r$ for all i and $\bar{e}_r = e_r$, the mean direction becomes: $\bar{g} = -w_r + W\bar{p}$.

Shared direction across training example gradients. We can observe in Equation (11) and Equation (13) that in case of static, repeating refusal token r , gradients across N examples will be concentrated around shared direction w_r . In case other term Wp does not cancel or bias w_r too much (for example, when p has large mass elsewhere), gradients concentrated around w_r will be low rank.

From gradients to gradient-induced activation updates. Activations are changed after parameter updates (training steps) minimizing aforementioned loss: the raw gradient g_i is not itself the realized activation change. Concretely, let $J_i := \nabla_\theta h_i(\theta)$ be its Jacobian. Then, $\nabla_\theta \ell_i^{\text{CE}} = J_i^\top g_i$. Under a small parameter update the realized change would be:

$$h_i(\theta - \eta \nabla_\theta \ell_i^{\text{CE}}) - h_i(\theta) \approx -\eta J_i J_i^\top g_i. \quad (14)$$

We define the *gradient-induced activation update* as the realized finite difference, taken with a negative sign for gradient-aligned comparison:

$$\mathbf{g}_i = - (h_i(\theta - \eta \nabla_{\theta} \ell_i^{\text{CE}}) - h_i(\theta)) \approx \eta J_i J_i^{\top} g_i \in \mathbb{R}^d. \quad (15)$$

Stacking these updates across N examples gives the *gradient-induced activation update matrix* $\mathcal{G}$:

$$\mathcal{G} = \begin{bmatrix} \mathbf{g}_1^{\top} \\ \vdots \\ \mathbf{g}_N^{\top} \end{bmatrix} \in \mathbb{R}^{N \times d}. \quad (16)$$

Training-time origin of refusal directions Assuming gradient descent over harmful prompts paired with refusal completions, we hypothesize that the refusal subspace $\mu_{\Delta} + \mathcal{V}(\Delta H_c)$ is approximated by the sign-reversed affine subspace of gradient-induced activation updates $-\mu_{\mathcal{G}} + \mathcal{V}(\mathcal{G}_c)$, where $\mathcal{G}_c$ denotes a mean-centered matrix $\mathcal{G}$

$$\mu_{\Delta} + \mathcal{V}(\Delta H_c) \approx -\mu_{\mathcal{G}} + \mathcal{V}(\mathcal{G}_c). \quad (17)$$

Concretely, we test the above prediction using these three diagnostics:

$$\cos(\mu_{\Delta}, \mu_{\mathcal{G}}) \approx -1, \quad \text{training steps move activations opposite to gradients,} \quad (18)$$

$$\mathcal{V}(\Delta H_c) \approx \mathcal{V}(\mathcal{G}_c), \quad \text{right singular vectors span similar subspaces,} \quad (19)$$

$$\text{sr}(\Delta H) \approx \text{sr}(\mathcal{G}) \quad \text{their stable ranks are very close.} \quad (20)$$

Stable rank of first-token target matrix. Let $E_r \in \mathbb{R}^{N \times |V|}$ be the one-hot matrix of refusal first tokens $r_1, \dots, r_N$. Across N examples, for a token $k \in V$, let $c_k = \#\{i : r_i = k\}$ denote its count, $f_k = c_k/N$ denote its empirical frequency, and let $m = |\{k : c_k > 0\}|$ be the number of observed first-token buckets. The following lemma, whose full statement, proof, and a corollary appear in Section E.1, shows that first-token concentration directly controls the stable rank of E_r .

Lemma 3.1 (Stable rank of refusal first-token targets). *The stable rank of E_r is*

$$\text{sr}(E_r) = \frac{N}{\max_k c_k} = \frac{1}{\max_k f_k}. \quad (21)$$

In particular, $\text{sr}(E_r) = 1$ when all examples share the same first refusal token. For a fixed number m of observed first-token buckets, $\text{sr}(E_r)$ is maximized by balanced counts across the buckets

$$\max \text{sr}(E_r) = \frac{N}{\lceil N/m \rceil}. \quad (22)$$

Thus, a repeated refusal-completion first token, such as the token beginning ‘‘I’m sorry,’’ minimizes $\text{sr}(E_r)$, while balanced first-token frequencies increase it. Since the last-layer first-token gradient matrix is $G_H^{\text{CE}} = (P - E_r)W^{\top}$, this gives a testable prediction: when the prediction term P and unembedding W do not collapse this target-rank signal, we can expect that increasing refusal first-token diversity can increase the stable rank of gradients G and their corresponding gradient-induced activation updates $\mathcal{G}$. We therefore report $\text{sr}(E_r)$, $\text{sr}(P - E_r)$, and $\text{sr}((P - E_r)W^{\top})$ separately to measure P and $W^{\top}$ preserve or distort the first-token target-matrix signal.

Stable rank distortion across layers: an idealized case. Assume in idealized case that for some corresponding Jacobian $J^{(l)} = \frac{\partial h^{(l)}}{\partial h^{(l-1)}} \in \mathbb{R}^{d \times d}$, gradients at previous layer $l - 1$ satisfy

$$G^{(l-1)} = G^{(l)} J^{(l)}, \quad (23)$$

while that Jacobian is shared across training examples. Assume also that the Jacobian is well-conditioned, i.e. its condition number $\kappa(J^{(l)}) = \|J^{(l)}\|_2 \|(J^{(l)})^{-1}\|_2$ is small. Under these assumptions, an inequality below bounds how much stable rank can change throughout remaining layers $1 \leq l \leq L - 1$ under this layerwise map; see Section E.2 for the proof.

Theorem 3.2 (Bounded stable rank distortion under a well-conditioned factor). *Let $A \in \mathbb{R}^{N \times n}$, let $B \in \mathbb{R}^{n \times n}$ be invertible, and let $\kappa(B) = \|B\|_2 \|B^{-1}\|_2$ be a condition number. Then*

$$\frac{1}{\kappa(B)^2} \text{sr}(A) \leq \text{sr}(AB) \leq \kappa(B)^2 \text{sr}(A). \quad (24)$$

Applying Theorem 3.2 to Equation (23) gives

$$\frac{1}{\kappa(J^{(l)})^2} \text{sr}(G^{(l)}) \leq \text{sr}(G^{(l-1)}) \leq \kappa(J^{(l)})^2 \text{sr}(G^{(l)}). \quad (25)$$

Iterating from layer L to layer l yields

$$\frac{\text{sr}(G^{(L)})}{\prod_{k=l+1}^L \kappa(J^{(k)})^2} \leq \text{sr}(G^{(l)}) \leq \left(\prod_{k=l+1}^L \kappa(J^{(k)})^2 \right) \text{sr}(G^{(L)}). \quad (26)$$

Assuming in our case $G^{(L)} = G_H^{\text{CE}} = (P - E_r)W^\top$ as per Equation (12), we obtain the layerwise prediction for l -layer gradients $G^{(l)}$, bounded by functions of $G^{(L)}$ from both sides:

$$\frac{\text{sr}((P - E_r)W^\top)}{\prod_{k=l+1}^L \kappa(J^{(k)})^2} \leq \text{sr}(G^{(l)}) \leq \left(\prod_{k=l+1}^L \kappa(J^{(k)})^2 \right) \text{sr}((P - E_r)W^\top). \quad (27)$$

This gives another testable prediction: under our idealized conditions, stable rank increases of last-layer's gradients $G^{(L)}$ can also propagate back, increasing stable rank of layer- l gradients $G^{(l)}$. The next parts refine the idealized condition above by introducing a data-dependent variant of a shared factor and data-dependent stable rank transfer factors in order to construct tighter bounds.

Definition 3.3 (Stable rank transfer factor). For nonzero gradient matrices $G^{(L)}$ and $G^{(l)}$ computed for the same N examples at layers L and l respectively, their stable rank transfer factor is the ratio of their stable ranks:

$$\tau^{L \rightarrow l} := \frac{\text{sr}(G^{(l)})}{\text{sr}(G^{(L)})}. \quad (28)$$

Thus, $\tau^{L \rightarrow l} > 1$ denotes stable rank amplification, while $\tau^{L \rightarrow l} < 1$ denotes stable rank contraction between the selected layers. We similarly define transfer factors for gradient-induced updates $\mathcal{G}$.

To refine impractical assumptions of Equation (23), instead of relying on an idealized, shared Jacobian, we introduce an appropriate approximation, or a shared cross-layer relation map.

Definition 3.4 (Top- k source-subspace shared map). For nonzero gradient matrices $G^{(L)}$ and $G^{(l)}$ computed for the same N examples at layers L and l respectively, a top- k source-subspace shared map $\hat{J}_k^{L \rightarrow l}$ is a shared data-dependent low-rank linear approximation for the observed cross-layer relation between gradients fitted on N examples and restricted to a top- k principal subspace of $G^{(L)}$. It is defined as follows:

$$\hat{J}_k^{L \rightarrow l} = V_k^{(L)} \Gamma_k^{L \rightarrow l}, \quad (29)$$

$$\hat{G}_k^{(l)} = G^{(L)} \hat{J}_k^{L \rightarrow l}. \quad (30)$$

where matrix $V_k^{(L)} \in \mathbb{R}^{d \times k}$ denotes top k right singular vectors of $G^{(L)}$, therefore $\text{rank}(\hat{J}_k^{L \rightarrow l}) \leq k$. $\Gamma_k^{L \rightarrow l} \in \mathbb{R}^{k \times d}$ is a coefficient matrix estimated from N gradients. $\hat{G}_k^{(l)}$ is the layer- l gradient matrix implied by the top- k source-subspace shared map $\hat{J}_k^{L \rightarrow l}$.

For a nonzero $\hat{G}_k^{(l)}$, the map-implied stable rank transfer factor is the stable rank transfer factor our shared map implies on $G^{(L)}$:

$$\hat{\tau}_k^{L \rightarrow l} = \frac{\text{sr}(\hat{G}_k^{(l)})}{\text{sr}(G^{(L)})} \quad (31)$$

Next, we account of errors resulting from the above approximation.

Definition 3.5 (Relative transfer factor fit error). For the observed $\tau^{L \rightarrow l}$ stable rank transfer factor, a shared top- k source-subspace map $\hat{J}_k^{L \rightarrow l}$, and a stable rank transfer factor $\hat{\tau}_k^{L \rightarrow l}$ implied by it, define a relative transfer factor error of $\hat{J}_k^{L \rightarrow l}$ as follows:

$$\epsilon_{\tau,k}^{L \rightarrow l} = \frac{|\hat{\tau}_k^{L \rightarrow l} - \tau^{L \rightarrow l}|}{\tau^{L \rightarrow l}}. \quad (32)$$

Theorem 3.2 gives a worst-case condition-number bound under an idealized shared linear factor. Lemma 3.6 instead shows how an upper bound on the transfer-factor fit error yields a data-dependent two-sided bound on the stable rank of $G^{(l)}$ (see Lemma E.5 for the proof); we will later estimate that upper bound empirically.

Lemma 3.6 (Stable rank bounds from transfer factor fit error). *Let $G^{(L)}, G^{(l)} \in \mathbb{R}^{N \times d}$ be nonzero gradient matrices, and let $\hat{J}_k^{L \rightarrow l}$ be a top- k source-subspace shared map with fitted output*

$$\hat{G}_k^{(l)} = G^{(L)} \hat{J}_k^{L \rightarrow l}.$$

Suppose that the relative transfer-factor fit error is bounded by some ϵ_{max} :

$$\epsilon_{\tau,k}^{L \rightarrow l} \leq \epsilon_{max} < 1.$$

Then

$$\frac{\hat{\tau}_k^{L \rightarrow l}}{1 + \epsilon_{max}} \text{sr}\left(G^{(L)}\right) \leq \text{sr}\left(G^{(l)}\right) \leq \frac{\hat{\tau}_k^{L \rightarrow l}}{1 - \epsilon_{max}} \text{sr}\left(G^{(L)}\right).$$

When ϵ_{max} is obtained by measuring the fit error on the same N examples used to estimate $\hat{J}_k^{L \rightarrow l}$, Lemma 3.6 is an in-sample fit certificate rather than an independent prediction of $\text{sr}(G^{(l)})$. It becomes a predictive propagation bound only when an independently established error bound ϵ_{max} is assumed to continue holding for additional source gradient matrices. Moreover, agreement at the level of stable rank does not imply matrix-level agreement; we evaluate the latter separately using relative Frobenius error, relative spectral error, and row-wise cosine similarity.

4 Empirical validation

4.1 Experimental setup

Target models. We perform measurements against allenai/OLMo-2-0425-1B-Instruct, an instruction-tuned OLMo 2 1B model with 16 transformer blocks, 16 attention heads, and hidden dimension $d = 2048$. For the training-stage analysis in Section 4.2, we evaluate its additional checkpoints: allenai/OLMo-2-0425-1B (olmo_base), allenai/OLMo-2-0425-1B-SFT (olmo_sft), allenai/OLMo-2-0425-1B-DPO (olmo_dpo), allenai/OLMo-2-0425-1B-RLVR1 (olmo_rlvr), and allenai/OLMo-2-0425-1B-Instruct (olmo_inst) [24].

Target datasets. We use AdvBench (adv), WildJailbreak evaluation and training splits (wld, wld_tr), and XSTest-Response subsets (xst_hr, xst_hr_jb, xst_rf, xst_rf_jb) as target prompt sets. Our frozen-model ablation analyses on these datasets use layer 7 for refusal vector estimation. For controlled fine-tuning, we use CAMEL chemistry dataset camel-ai/chemistry [12] by pairing its prompts with refusals sampled from a predefined set of refusals we manually collect; we use layer 8 to estimate refusal vector.

Metrics. We use refusal (harm) rate, or a proportion of refused (or harmful) prompts to all prompts as downstream metrics, using allenai/wildguard as a judge LLM for refusal (harm) scores; refusal delta (change under attack) is a difference between pre-intervention and post-intervention refusal rates. To test the affine-subspace prediction in Equation (17), we compare three quantities for a benign-centered residual matrix ΔH paired with gradient-induced activation updates $\mathcal{G}$ (or, similarly, paired with raw gradients G): (1) cosine similarity between their ΔH and $\mathcal{G}$ mean vectors; (2) stable rank on the uncentered ΔH and $\mathcal{G}$; (3) average principal-angle overlap between their centered top- k right-singular subspaces. We report stable ranks and transfer-factor summary statistics across conditions. Shared-map fits are evaluated using source-subspace energy, relative Frobenius and spectral errors, row-wise cosine similarity, and relative transfer-factor fit error, with identity and scalar-identity baselines. Full definitions appear in Section F.

4.2 Refusal direction across released post-training checkpoints

We first ask when the difference-in-means vector becomes a functional refusal direction. Using 60 AdvBench prompts, we estimate refusal vectors ρ at each layer for the OLMo checkpoints and measure refusal rates before/after ablation, refusal vector norm, cosine similarity to the final olmo_inst

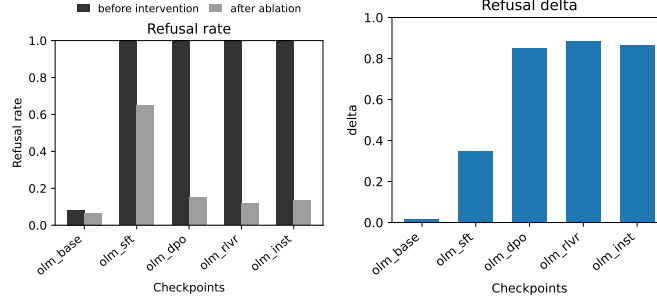


Figure 2: Descriptive comparison of released OLMo checkpoints along Base $\rightarrow$ SFT $\rightarrow$ DPO $\rightarrow$ RLVR1 $\rightarrow$ Instruct. The base checkpoint is not attack-effective under this estimator. A functional refusal direction is present by SFT, while post-SFT checkpoints exhibit larger refusal deltas than SFT.

direction, and stable rank of ΔH . Figure 2 shows that the base checkpoint has little refusal behavior and a negligible refusal delta under this estimator, whereas all four post-base checkpoints refuse nearly all selected prompts before intervention. A functional refusal-mediating direction is present at the SFT checkpoint, but its refusal delta is substantially smaller than those of the DPO, RLVR1, and Instruct checkpoints. Because the checkpoints differ in their objectives, data, and optimization histories, this comparison does not isolate the causal effect of any individual post-training stage. Figure 13 shows that the refusal vectors estimated at SFT and the later checkpoints remain highly aligned with the final `olmo_inst` direction across layers, while their stable rank and vector norm profiles largely overlap. Thus, the released post-SFT checkpoints approximately preserve the refusal geometry even though SFT is less attack-effective in our settings. Finally, Figure 12 shows that, for `olmo_inst`, directions estimated at layers 7–10 produce the largest refusal deltas.

4.3 Gradient-induced activation updates match benign-centered refusal residuals

We next test the affine-subspace prediction in Equation (17): benign-centered refusal residuals ΔH should share their mean direction, principal subspace, and stable rank with gradient-induced activation updates $\mathcal{G}$, up to the sign induced by gradient descent. Using the target datasets introduced above, we compute ΔH and $\mathcal{G}$ for three target-completion types: (1) model-generated refusal completions, (2) post-ablation completions, and (3) fixed pseudorandom English-token completions. For each dataset and output type, we compare ΔH and $\mathcal{G}$ using mean-direction cosine similarity, average principal-angle overlap between centered right-singular subspaces, and stable rank. We make similar comparisons for ΔH and raw gradients G as well.

The left column of Figure 3 shows that refusal-based means are negatively aligned with the difference-in-means refusal vector, while ablated-jailbreak-based and random-English-based means are positively aligned with it. Thus, refusal and non-refusal targets move along approximately the same refusal axis but in opposite directions, supporting Equation (18). This effect appears primarily for the refusal directions estimated at middle layer 7 and partially from the last layer 15. The same column also shows average principal-angle overlap between $\mathcal{V}(\Delta H_c)$ and $\mathcal{V}(\mathcal{G}_c)$. At layer 15, refusal-target updates subspace overlap refusal subspace more than both random-English and ablated-jailbreak controls gradient update subspaces for every evaluated dataset, indicating stronger last-layer agreement between direction subspaces of refusal gradient updates $\mathcal{V}(\mathcal{G}_c)$ and refusal subspace $\mathcal{V}(\Delta H_c)$. At layer 7, overall subspace overlap is lower, and the overlap appears to be more strongly associated with the dataset than with the completion type: the three target types within a dataset are relatively close, while the evaluation split of WildJailbreak (`wld`) has lower overlap than its training split (`wld_tr`) and most other datasets.

The right column of Figure 3 shows that, for refusal targets, the stable ranks of $\mathcal{G}$ and ΔH are close across datasets for both layer 7 and layer 15. The observed stable ranks are below the sample size $N = 60$, around 1–3, indicating genuine spectral concentration and excluding trivial rank cap from insufficient datapoints. We observe the same qualitative relationship across most transformer layers.

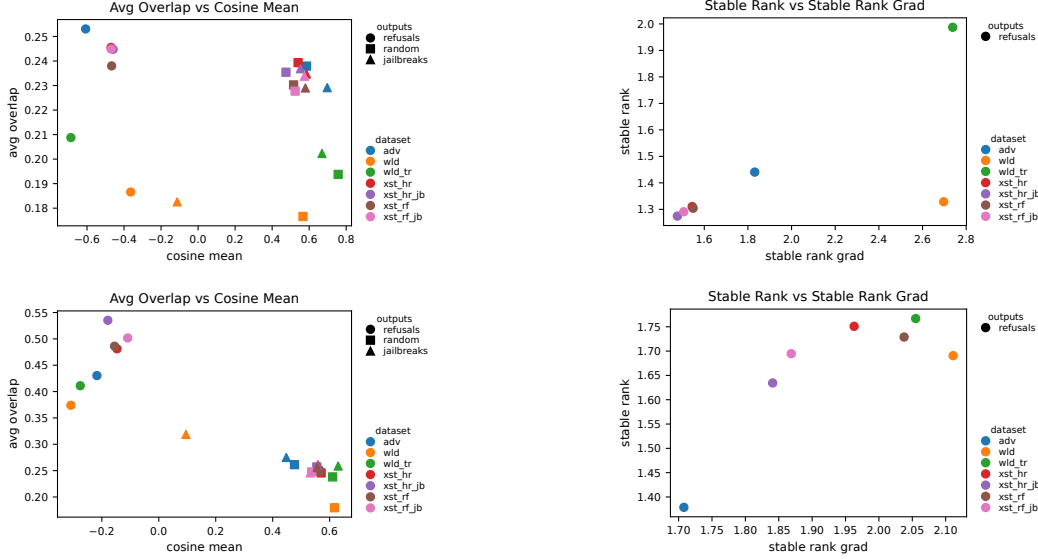


Figure 3: Gradient-induced activation updates explain benign-centered refusal residuals. Left column: mean alignment and average principal-angle overlap. Right column: stable rank comparison. Top row: layer 7; bottom row: layer 15.

Together, these three diagnostics support the main empirical implication of Equation (17): signed mean-direction alignment, partial and layer-dependent principal-subspace overlap, and closely matched stable ranks. They support an approximate affine-subspace relationship.

As a complementary uncentered diagnostic, at the refusal-estimating layer 7, the leading right-singular axis of $\mathcal{G}$ has mean absolute cosine similarity 0.56 with the refusal direction and 0.96 with the mean of $\mathcal{G}$ across the seven evaluated datasets. Thus, the dominant uncentered gradient-update axis of $\mathcal{G}$ is nearly collinear, up to sign, with the mean gradient-update vector of $\mathcal{G}$ and moderately aligned with the refusal direction. Together with the low stable rank of $\mathcal{G}$, this makes the mean-direction and leading-axis descriptions of its update geometry nearly equivalent in this setting and motivates a direct comparison between difference-in-means refusal ablation and ablation along the top largest principal directions.

4.4 Refusal first-token diversity increases gradient stable ranks

We next test on a frozen model whether varying refusal first tokens on a fixed prompt set changes the stable ranks of first-token gradients from the last layer to refusal-mediating middle layers. Using the WildJailbreak training split, we fix 80 harmful prompts whose model-generated refusals begin with the common first token “I”. For each $m \in \{1, 2, 4, 6, 8, 10, 12, 14, 16\}$, we pair each prompt with a refusal sampled independently with replacement from a pool of m short phrases with distinct first tokens. For each setting, we compute E_r , P , and $P - E_r$, the last-layer activation gradient matrix $(P - E_r)W^\top$, and the layerwise gradient matrices $G^{(l)}$ for the first-token cross-entropy loss. We use raw gradients in this experiment to test the cross-layer rank-transfer analysis from Section 3; gradient-induced activation updates are considered in the next subsection.

Figure 4 shows that all target-space quantities and gradients across layers 7-15 increase with the number of refusals: gradient stable rank increase is bounded and quickly saturates for the last layer due to constant P , but amplified throughout mid-layers on a backward pass. Across the nine settings, the observed layerwise transfer factor mean $\pm$ sample standard deviation are $\tau^{15 \rightarrow 7} = 2.270 \pm 0.175$, with range $[1.925, 2.469]$. Thus, every evaluated setting exhibits approximately twofold stable-rank amplification from layer 15 to layer 7. By contrast, the end-to-end $E_r \rightarrow G^{(7)}$ ratio is more variable, with mean 0.942 ± 0.486 and range $[0.504, 1.942]$.

For the $m = 16$ setting, we additionally fit the top- k source-subspace shared map from $G^{(15)}$ to $G^{(7)}$ on the same 80 examples and sweep k . At $k = 20$, the selected source subspace captures 99.8%

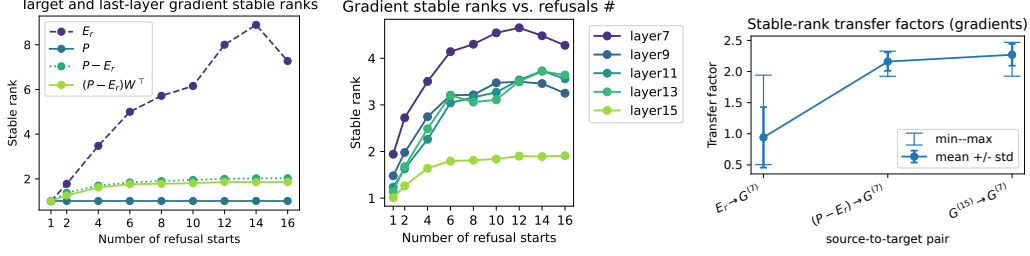


Figure 4: Stable ranks and transfer under refusal-start diversification. Left: stable ranks of the target-space quantities and the resulting last-layer gradient. Middle: stable ranks of the first-token gradients at selected middle-to-late layers. Right: across-setting layerwise stable rank transfer factor mean $\pm$ std and min/max. The stable ranks of E_r , $P - E_r$, and $G^{(l)}$ increase overall. The $G^{(15)} \rightarrow G^{(7)}$ transfer factor varies substantially less than the end-to-end $E_r \rightarrow G^{(7)}$ ratio.

of the source-gradient energy; the relative Frobenius and spectral errors are 0.594 and 0.316, the mean row-wise cosine similarity is 0.806, and the relative transfer-factor fit error is 0.343. The fitted map yields lower reconstruction and transfer-factor errors and higher row-wise cosine similarity than the identity and best scalar-identity baselines. This provides evidence that the observed in-sample cross-layer relation admits a meaningful but incomplete shared linear approximation. The complete rank sweep and reduced-map spectrum are reported in Figures 14 and 15.

4.5 Refusal first-token diversity increases residual rank and weakens single-vector ablation

We next test on a frozen model whether increasing the diversity of first tokens across refusal phrases (through varying selection of prompts in the dataset) increases the effective rank of benign-centered refusal residuals and gradient-induced activations updates. Using the WildJailbreak training split, we select harmful prompts that the target model refuses, keeping the baseline refusal rate fixed across subsets. We construct nine 80-example subsets drawing from $m \in \{1, 2, 4, 6, 8, 10, 12, 14, 16\}$ refusal first-token buckets. For each subset, we compute benign-centered refusal residuals ΔH , raw first-token activation gradients $G^{(l)}$, gradient-induced activation updates $\mathcal{G}^{(l)}$, and refusal and harmfulness scores before and after difference-in-means ablation. Both $G^{(l)}$ and $\mathcal{G}^{(l)}$ are computed using the original WildJailbreak response paired with each selected prompt. We then plot the ablation score deltas against the stable rank of ΔH . We select layer 7 as the earliest layer in the range of effective layers.

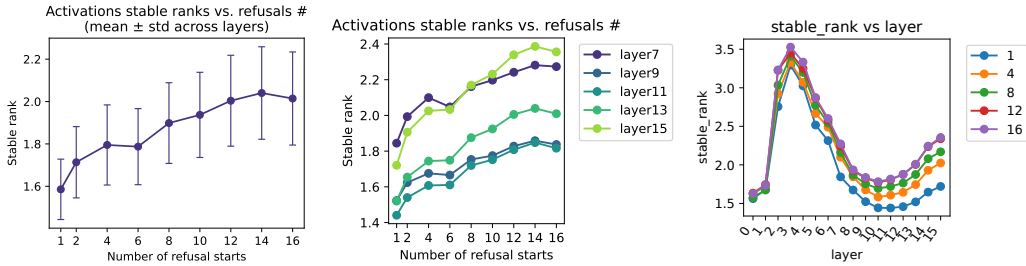


Figure 5: Stable rank of benign-centered refusal residuals ΔH . Left plot averaged ranks over layers 7-15. Increasing refusal first-token diversity raises $\text{sr}(\Delta H)$ in middle-to-late layers.

Figure 5 shows that increasing the number of refusal first-token buckets raises ΔH stable rank. The effect is visible across middle-to-late layers, while the residual representation at earlier layers retain their stable ranks. Figure 17 shows similar trend for gradient-induced activation updates, but with saturation at the end. As the refusal first tokens become more diverse, the stable rank of $\mathcal{G}$ also increases across most layers, however it saturates at 12-16 refusal starts.

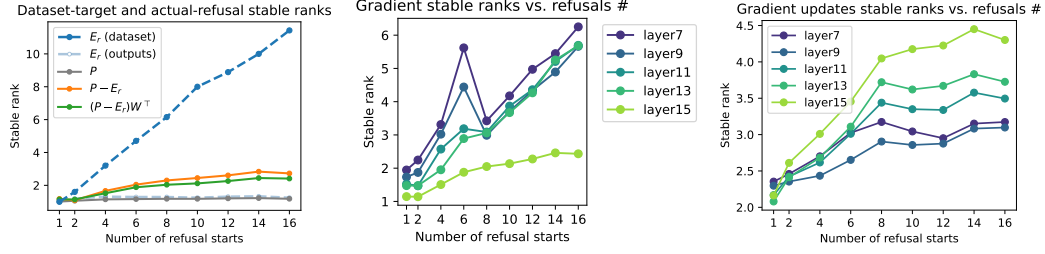


Figure 6: Stable rank propagation under refusal-start diversification by prompt selection. Left: stable ranks of the first-token target matrix E_r , prediction matrix P , logit-gradient matrix $P - E_r$, and last-layer activation-gradient matrix $(P - E_r)W^\top$. Middle: stable ranks of the first-token gradients $G^{(l)}$. Right: stable ranks of the corresponding gradient-induced activation updates $\mathcal{G}^{(l)}$. The target-space and layerwise stable ranks increase overall as refusal-start support broadens.

Figure 6 shows the same overall rank increase from the dataset-target quantities through the raw gradients $G^{(l)}$ and the gradient-induced activation updates $\mathcal{G}^{(l)}$. Notably, the dataset-derived E_r rank rises sharply across the prompt subsets, whereas the model’s actual output-refusal-start E_r rank remains much lower. Thus, the dataset responses contain substantially more first-token diversity than the frozen model realizes in its own completions.

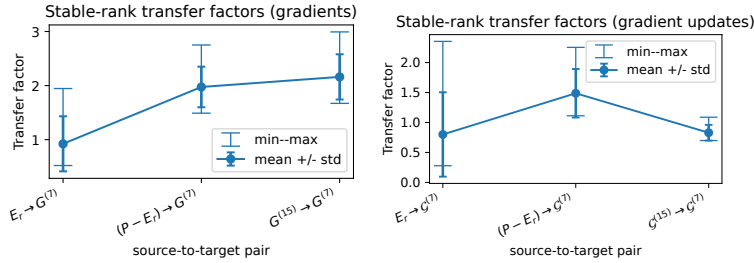


Figure 7: Variation of stable rank transfer factor for raw gradients and gradient-induced activation updates across experimental settings. Left: raw gradient transfer factor. Right: gradient-induced activation deltas transfer factor. Raw gradients consistently amplify stable rank between layers 15 and 7, whereas the gradient update matrices undergo mild stable-rank contraction on average.

Across the nine settings, raw gradient transfer factor is $\tau_G^{15 \rightarrow 7} = 2.161 \pm 0.418$, reported as mean $\pm$ sample standard deviation, with range $[1.671, 2.993]$. Every evaluated setting therefore exhibits stable rank amplification from $G^{(15)}$ to $G^{(7)}$, by approximately a factor of two on average, although the amplification magnitude varies across subsets. For gradient-induced activation updates, the corresponding factor is $\tau_{\mathcal{G}}^{15 \rightarrow 7} = 0.829 \pm 0.132$, with range $[0.698, 1.088]$ indicating mild stable-rank contraction on average. The layer-15-to-layer-7 update transfer is more stable across subsets than the full target-to-middle-layer ratios. The complete in-sample shared-map rank sweeps and reduced-map spectra for G and $\mathcal{G}$ are reported in Figures 18 to 21.

We then test whether higher stable rank of ΔH corresponds to weaker single-vector ablation. Figure 8 reports harmfulness scores before and after ablation, and plots the harmfulness delta against $\text{sr}(\Delta H)$. Subsets with higher-rank ΔH have smaller harmfulness deltas, suggesting that a less one-dimensional residual subspace is harder to suppress with one vector. Figure 9 shows, similarly, that the refusal delta slightly decreases as $\text{sr}(\Delta H)$ increases, indicating that higher-rank residuals ΔH weaken the effect of single-vector refusal ablation.

Overall, this fixed-model subset analysis supports the rank mechanism: increasing refusal first-token diversity raises the effective rank of both ΔH and $\mathcal{G}$, while higher activation stable rank is associated with smaller refusal and harmfulness deltas under single-vector ablation.

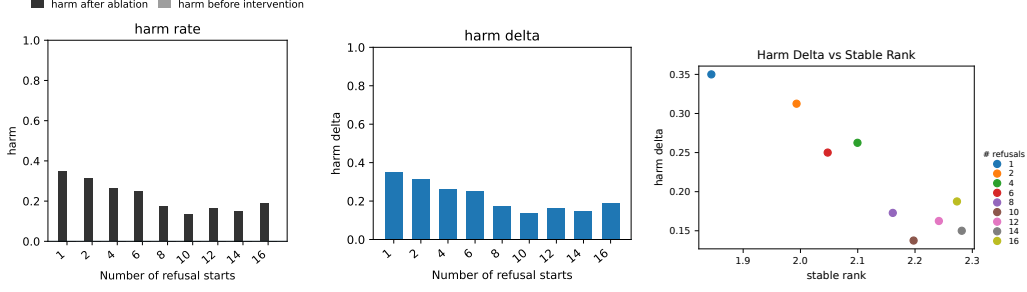


Figure 8: Harmfulness score decomposition and harmfulness delta versus $\text{sr}(\Delta H)$. Higher refusal residuals stable rank is associated with smaller harmfulness deltas under difference-in-means ablation.

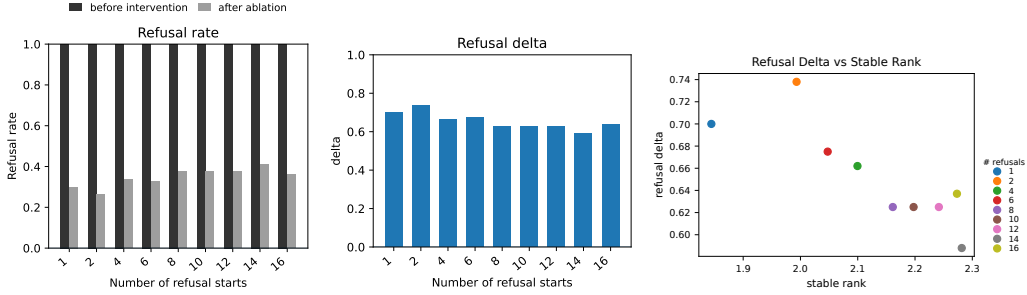


Figure 9: Refusal score decomposition and refusal delta versus $\text{sr}(\Delta H)$. Higher activation stable rank is associated with smaller refusal deltas, consistent with weaker single-vector ablation.

4.6 Diverse-refusal fine-tuning raises refusal residuals rank and weakens ablation

We next test whether the rank–vulnerability pattern from Section 4.5 also appears when simulating refusal training on the prompts are initially benign and not refused. The primary prediction is that increasing refusal first-token diversity should raise refusal residuals stable rank and reduce the refusal delta after difference-in-means single vector ablation. Attacking by ablating more principal directions is possible only if this does not degrade performance and the directions generalize across the prompts (we leave multiple-vector ablation attacks out of scope). We keep the prompt set fixed and vary only the generated refusal-style completions by sampling 40 new-topic prompts from *camel-ai/chemistry* and pairing them with refusal completions sampled from $m \in \{1, 4, 8, 12, 16\}$ distinct first-token buckets with balanced counts within each dataset, which directly manipulates the stable rank of the refusal first-token matrix E_r . We fine-tune OLMo-2-0425-1B-Instruct separately on these datasets for each m and select each time the earliest checkpoint that reaches or exceeds refusal rate threshold 0.6, so that later measurements are done under close base refusal rates. For intervention, we select layer 8 as the earliest layer in the range of effective layers.

Additionally, for $m = 8$ and $m = 16$, we track refusal rates across optimization steps. Figure 22 shows that the ablation effect is largest in a narrow optimization window. This suggests that, although the number of optimization steps for each m is different, fixing the resulting checkpoints at first-reached baseline rates of 0.6 compares largest attack affects.

At the matched checkpoints, we first measure the stable rank of the resulting residuals ΔH . Figure 10 shows that stable rank of ΔH increases with the number of refusal first-token buckets with the largest effect across mid-to-late layers. This supports the rank mechanism: more diverse refusal starts induce less concentrated activation changes.

We then evaluate whether this rank increase weakens single-vector ablation. Figure 11 shows refusal scores before and after ablation at the matched checkpoints as well as stable rank of ΔH plotted against refusal deltas. Figure 23 shows that, similar to the previous sections, with diverse fine-tuning, stable rank of target quantities as well as gradient-induced update deltas increase on average,

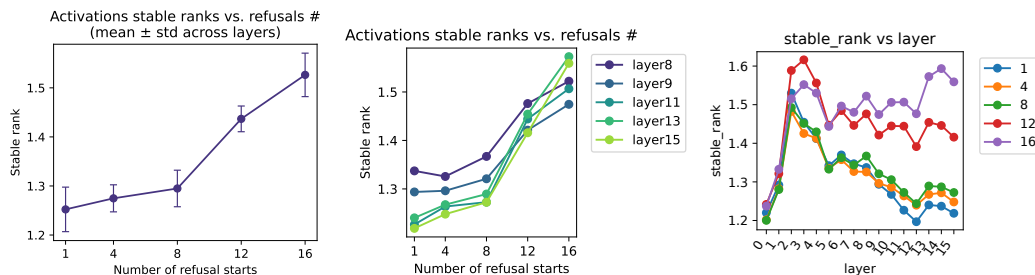


Figure 10: Stable rank of refusal residuals under diverse-refusal fine-tuning. Left plot averages ranks across layers 8-15. Stable ranks across layers and their per-layer averages increase with the number of refusal first-token buckets with the clearest separation in middle-to-late layers.

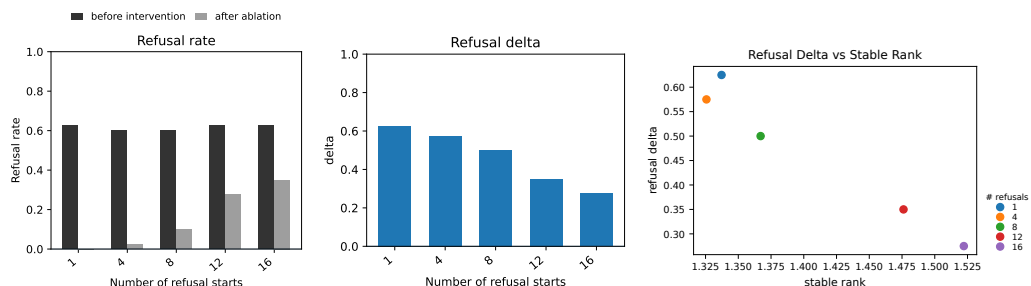


Figure 11: Refusal ablation at checkpoints matched by baseline refusal score. Left and middle plots: diverse refusals yield smaller deltas under difference-in-means ablation. Right plot: diverse refusal starts are associated with higher refusal residuals stable rank and smaller refusal deltas

undergoing mild stable rank expansion from layer 15 to layer 7. Finally, rightmost Figure 11 directly relates geometry to attackability. More diverse refusal starts are associated with higher refusal residuals stable rank and smaller refusal deltas under ablation. This suggests that increasing refusal first-token diversity weakens the single dominant refusal vector by spreading the fine-tuning update over more effective directions.

Overall, this controlled fine-tuning experiment suggests the rank–vulnerability pattern can hold broadly. Even on non-harmful chemistry prompts, concentrating refusal first tokens produces lower-rank refusal residuals and stronger single-vector ablation, while diversifying refusal starts raises the refusal residuals stable ranks and reduces the ablation effect at matched baseline refusal score.

5 Conclusion

As AI systems enter dual-use domains, robustness failures can be catastrophic and unevenly impact at-risk communities. Our work takes a step toward understanding AI vulnerabilities at the root cause and designing defenses from first principles.

We asked why refusal behavior can concentrate in a low-dimensional activation-space mechanism that is shared across many harmful categories and can be effectively suppressed. In OLMo-2-0425-1B-Instruct, we find that the refusal directions reflect refusal safety-training: cross-entropy gradients from the first tokens of refusal completions induce activation updates whose mean and principal directions align with refusal direction and refusal subspace. This connects the empirical difference-in-means refusal vector to the training updates that create it. This mechanism suggests a simple source of brittleness exist: concentrated refusal first tokens induce low-stable-rank activation changes, making refusal behavior easier to suppress with a single-vector refusal ablation. Across fixed-model and controlled fine-tuning experiments, increasing refusal first-token diversity raises stable rank

of the activation-change matrix and is associated with smaller refusal deltas. At the same time, refusal-completion diversity alone is unlikely a complete defense.

Overall, our results suggest that refusal robustness to ablation is partly shaped by the geometry of refusal training updates. Token-level properties of the training completions, especially refusal-prefix concentration and first-token frequency, can shape the dimensionality of the learned refusal mechanism. Understanding this link between data, gradients, and activation geometry may be a step toward principled mechanistic understanding defenses against jailbreaks. Finally, our controlled fine-tuning experiments on non-harmful datasets motivate study whether similar mechanisms hold for a broader set of features than for refusal or safety-related features.

References

- [1] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. 2nd. Cambridge: Cambridge University Press, 2012. ISBN: 978-0521548236.
- [2] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. *Plug and Play Language Models: A Simple Approach to Controlled Text Generation*. 2020. arXiv: [1912.02164](https://arxiv.org/abs/1912.02164) [cs.CL]. URL: <https://arxiv.org/abs/1912.02164>.
- [3] Vardan Papyan, XY Han, and David L. Donoho. “Prevalence of neural collapse during the terminal phase of deep learning training”. In: *Proceedings of the National Academy of Sciences* 117.40 (2020), pp. 24652–24663.
- [4] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. *Attention Is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth*. 2021. arXiv: [2103.03404](https://arxiv.org/abs/2103.03404) [cs.LG]. URL: <https://arxiv.org/abs/2103.03404>.
- [5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. *Constitutional AI: Harmlessness from AI Feedback*. 2022. arXiv: [2212.08073](https://arxiv.org/abs/2212.08073) [cs.CL]. URL: <https://arxiv.org/abs/2212.08073>.
- [6] 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. arXiv: [2203.02155](https://arxiv.org/abs/2203.02155) [cs.CL]. URL: <https://arxiv.org/abs/2203.02155>.
- [7] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. *Red Teaming Language Models with Language Models*. 2022. arXiv: [2202.03286](https://arxiv.org/abs/2202.03286) [cs.CL]. URL: <https://arxiv.org/abs/2202.03286>.
- [8] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. *Extracting Latent Steering Vectors from Pretrained Language Models*. 2022. arXiv: [2205.05124](https://arxiv.org/abs/2205.05124) [cs.CL]. URL: <https://arxiv.org/abs/2205.05124>.
- [9] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O’Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tatum Collins, Tim Fisk, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. 2023. arXiv: [2307.03718](https://arxiv.org/abs/2307.03718) [cs.CY]. URL: <https://arxiv.org/abs/2307.03718>.
- [10] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. *Jailbreaking Black Box Large Language Models in Twenty Queries*. 2023. arXiv: [2310.08419](https://arxiv.org/abs/2310.08419) [cs.LG]. URL: <https://arxiv.org/abs/2310.08419>.
- [11] Anjali Gopal, Nathan Helm-Burger, Lennart Justen, Emily H. Soice, Tiffany Tzeng, Geetha Jeyapragasan, Simon Grimm, Benjamin Mueller, and Kevin M. Esvelt. *Will Releasing the Weights of Future Large Language Models Grant Widespread Access to Pandemic Agents?* 2023. arXiv: [2310.18233](https://arxiv.org/abs/2310.18233) [cs.AI]. URL: <https://arxiv.org/abs/2310.18233>.
- [12] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. *CAMEL: Communicative Agents for “Mind” Exploration of Large Scale Language Model Society*. 2023. arXiv: [2303.17760](https://arxiv.org/abs/2303.17760) [cs.AI]. URL: <https://arxiv.org/abs/2303.17760>.
- [13] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. *Tree of Attacks: Jailbreaking Black-Box LLMs Automatically*. 2023. arXiv: [2312.02119](https://arxiv.org/abs/2312.02119) [cs.LG]. URL: <https://arxiv.org/abs/2312.02119>.
- [14] Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. “Feature learning in deep classifiers through intermediate neural collapse”. In: *International conference on machine learning*. PMLR. 2023, pp. 28729–28745.
- [15] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. *Jailbroken: How Does LLM Safety Training Fail?* 2023. arXiv: [2307.02483](https://arxiv.org/abs/2307.02483) [cs.LG]. URL: <https://arxiv.org/abs/2307.02483>.

- [16] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. 2023. arXiv: [2307.15043](https://arxiv.org/abs/2307.15043) [cs.CL]. URL: <https://arxiv.org/abs/2307.15043>.
- [17] 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. arXiv: [2406.11717](https://arxiv.org/abs/2406.11717) [cs.LG]. URL: <https://arxiv.org/abs/2406.11717>.
- [18] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. *Stealing Part of a Production Language Model*. 2024. arXiv: [2403.06634](https://arxiv.org/abs/2403.06634) [cs.CR]. URL: <https://arxiv.org/abs/2403.06634>.
- [19] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. Tech. rep. RR-A2849-1. RAND Corporation, 2024. DOI: [10.7249/RR2849-1](https://www.rand.org/pubs/research_reports/RR2849-1.html). URL: https://www.rand.org/pubs/research_reports/RR2849-1.html.
- [20] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. *Steering Llama 2 via Contrastive Activation Addition*. 2024. arXiv: [2312.06681](https://arxiv.org/abs/2312.06681) [cs.CL]. URL: <https://arxiv.org/abs/2312.06681>.
- [21] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. *Safety Alignment Should Be Made More Than Just a Few Tokens Deep*. 2024. arXiv: [2406.05946](https://arxiv.org/abs/2406.05946) [cs.CR]. URL: <https://arxiv.org/abs/2406.05946>.
- [22] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. *Function Vectors in Large Language Models*. 2024. arXiv: [2310.15213](https://arxiv.org/abs/2310.15213) [cs.CL]. URL: <https://arxiv.org/abs/2310.15213>.
- [23] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. *Steering Language Models With Activation Engineering*. 2024. arXiv: [2308.10248](https://arxiv.org/abs/2308.10248) [cs.CL]. URL: <https://arxiv.org/abs/2308.10248>.
- [24] Allen Institute for AI. *OLMo-2-0425-1B-Instruct Model Card*. <https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct>. Hugging Face model card. Accessed: 2026-05-07. 2025.
- [25] Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. *Steering Large Language Model Activations in Sparse Spaces*. 2025. arXiv: [2503.00177](https://arxiv.org/abs/2503.00177) [cs.LG]. URL: <https://arxiv.org/abs/2503.00177>.
- [26] Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Mikhail Belkin. *Toward universal steering and monitoring of AI models*. 2025. arXiv: [2502.03708](https://arxiv.org/abs/2502.03708) [cs.CL]. URL: <https://arxiv.org/abs/2502.03708>.
- [27] Marie Davidsen Buhl, Ben Bucknall, and Tammy Masterson. *Emerging Practices in Frontier AI Safety Frameworks*. 2025. arXiv: [2503.04746](https://arxiv.org/abs/2503.04746) [cs.CY]. URL: <https://arxiv.org/abs/2503.04746>.
- [28] Jacob Dunefsky and Arman Cohan. *One-shot Optimized Steering Vectors Mediate Safety-relevant Behaviors in LLMs*. 2025. arXiv: [2502.18862](https://arxiv.org/abs/2502.18862) [cs.LG]. URL: <https://arxiv.org/abs/2502.18862>.
- [29] Frontier Model Forum. *Frontier Mitigations*. Tech. rep. Accessed: 2026-05-05. Frontier Model Forum, June 2025. URL: <https://www.frontiermodelforum.org/technical-reports/frontier-mitigations/>.
- [30] Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. *SAIF: A Sparse Autoencoder Framework for Interpreting and Steering Instruction Following of Language Models*. 2025. arXiv: [2502.11356](https://arxiv.org/abs/2502.11356) [cs.LG]. URL: <https://arxiv.org/abs/2502.11356>.
- [31] Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. *Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences*. 2025. arXiv: [2510.13900](https://arxiv.org/abs/2510.13900) [cs.CL]. URL: <https://arxiv.org/abs/2510.13900>.
- [32] Model Evaluation and Threat Research. *Common Elements of Frontier AI Safety Policies*. Accessed: 2026-05-05. 2025. URL: <https://metr.org/common-elements>.

- [33] Kyle O’Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. *Steering Language Model Refusal with Sparse Autoencoders*. 2025. arXiv: [2411.11296](https://arxiv.org/abs/2411.11296) [cs.LG]. URL: <https://arxiv.org/abs/2411.11296>.
- [34] OpenAI. *Preparing for Future AI Capabilities in Biology*. Accessed: 2026-05-05. June 2025. URL: <https://openai.com/index/preparing-for-future-ai-capabilities-in-biology/>.
- [35] Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. *The Hidden Dimensions of LLM Alignment: A Multi-Dimensional Analysis of Orthogonal Safety Directions*. 2025. arXiv: [2502.09674](https://arxiv.org/abs/2502.09674) [cs.CL]. URL: <https://arxiv.org/abs/2502.09674>.
- [36] 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*. 2025. arXiv: [2509.09708](https://arxiv.org/abs/2509.09708) [cs.CL]. URL: <https://arxiv.org/abs/2509.09708>.
- [37] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. *Open Problems in Mechanistic Interpretability*. 2025. DOI: [10.48550/arXiv.2501.16496](https://arxiv.org/abs/2501.16496). arXiv: [2501.16496](https://arxiv.org/abs/2501.16496) [cs.LG]. URL: <https://arxiv.org/abs/2501.16496>.
- [38] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. *Improving Instruction-Following in Language Models through Activation Steering*. 2025. arXiv: [2410.12877](https://arxiv.org/abs/2410.12877) [cs.CL]. URL: <https://arxiv.org/abs/2410.12877>.
- [39] Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. *The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence*. 2025. arXiv: [2502.17420](https://arxiv.org/abs/2502.17420) [cs.LG]. URL: <https://arxiv.org/abs/2502.17420>.
- [40] Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. *Robust LLM safeguarding via refusal feature adversarial training*. 2025. arXiv: [2409.20089](https://arxiv.org/abs/2409.20089) [cs.LG]. URL: <https://arxiv.org/abs/2409.20089>.
- [41] Anthropic. *Assessing Claude Mythos Preview’s Cybersecurity Capabilities*. Accessed: 2026-05-05. Apr. 2026. URL: <https://red.anthropic.com/2026/mythos-preview/>.
- [42] Yoshua Bengio, Stephen Clare, Carina Prunkl, et al. *International AI Safety Report 2026*. 2026. arXiv: [2602.21012](https://arxiv.org/abs/2602.21012) [cs.CY]. URL: <https://arxiv.org/abs/2602.21012>.
- [43] Taehun Cha, Daniel Beaglehole, Adityanarayanan Radhakrishnan, and Donghun Lee. *The Weight Gram Matrix Captures Sequential Feature Linearization in Deep Networks*. 2026. arXiv: [2605.06258](https://arxiv.org/abs/2605.06258) [cs.LG]. URL: <https://arxiv.org/abs/2605.06258>.
- [44] Damek Davis and Dmitriy Drusvyatskiy. *When do spectral gradient updates help in deep learning?* 2026. arXiv: [2512.04299](https://arxiv.org/abs/2512.04299) [cs.LG]. URL: <https://arxiv.org/abs/2512.04299>.

A Related work

A.1 Activation steering and ablation attacks

Activation steering and ablation attacks are a family of whitebox methods which directly act on model activations and runtime to alter its output in a desired way. Those methods are widely used to suppress refusals, bypass safety alignment, and to jailbreak open models. An early work [23] introduces activation engineering, an inference-time steering of model activation by adding a vector, computed from contrasting datasets. [17] introduced a difference-of-means jailbreak attack: the method works by estimating refusal vector as a difference between harmful and harmless prompts’ activation means and later removing projections on this vector to let model respond to prompts it was trained to refuse. [40] additionally confirms that the operation of refusal feature ablation approximates the worst-case perturbation of offsetting model safety. [28] find a steering vector by performing gradient descent from a single training example to mediate safety-relevant behavior. [26] offers a universal steering by adding eigenvectors extracted from Recursive Feature Machines (RFMs) trained on contrasting datasets.

A line of work introduces sparse autoencoder (SAE) based steering. O’Brien et al. identify and amplify SAE features mediating refusal [33]; Bayat et al. introduce sparse activation steering by selecting SAE features from contrastive prompts [25]; SAIF steers instruction-following behavior through instruction-relevant SAE latents [30].

Activation steering has been useful more broadly in non-security contexts, too: ActAdd and CAA use prompt-pair or contrastive activation differences to steer behaviors, such as topic, sentiment, factuality, or behavior [23], [20]; PPLM steers generation by using gradients from an attribute model to act on activations [2]; Subramani et al. extract steering vectors from frozen models with gradient descent [8]; Function Vectors and instruction-steering work extract task or instruction vectors for inference-time control [22], [38].

Our work complements steering attack approaches by explaining a notable safety-critical feature in activations, the refusal vector, and its low-dimensional structure using analytical methods.

A.2 Mechanistic interpretability, representation geometry, and safety-training dynamics

An adjacent body of work studies the representations learned by models and the training dynamics that produce them. [39] show that refusal representations span multiple orthogonal directions and can be described as a concept cone, motivating representational independence: interventions on different directions can mediate different behaviors without necessarily affecting others. Similarly, [35] argue that refusal is multi-dimensional: a dominant direction mediates refusal, while additional components correspond to distinct interpretable safety features. [37] survey open problems and research areas in mechanistic interpretability, including predicting which capabilities and mechanisms arise during training or fine-tuning. Complementing these empirical and conceptual accounts of multi-dimensional refusal, our work analytically studies the origin of such orthogonal refusal directions and links them to training-time refusal gradients. We further provide evidence that concentrated refusal datasets can produce lower-rank gradient-induced activation updates, yielding low-stable-rank activation changes and nearly one-dimensional refusal representations; this structure, in turn, predicts larger effectiveness of the ablation attack.

A broader body of work studies concentrations as well as rank and dimensionality collapse. [31] demonstrate that fine-tuning on a narrow domain leaves identifiable traces of the training objective in model activations, suggesting that semantically narrow fine-tuning can induce readable activation-space biases. [21] find that the largest safety fine-tuning updates fall on the first assistant tokens, and that constraining updates to be distributed more evenly across output tokens improves robustness to fine-tuning-based jailbreaks. [44] identify conditions under which spectral optimization is preferred to Euclidean gradient descent and show that transformer activations can remain low-stable-rank during training. [4] prove that pure self-attention can converge toward rank-one token-uniform representations without skip connections or MLPs. [3] describe terminal-training neural collapse, in which final-layer class features collapse to class means with highly symmetric geometry, while [14] extend this picture to intermediate layers, showing that deeper representations increasingly reduce within-class variance relative to between-class variance and align class-mean subspaces with dominant weight directions. Unlike these broader investigations of rank collapse and neural collapse,

our work focuses on safety post-training: we study a mechanism in which concentration of refusal first tokens makes the target matrix low stable rank, inducing low-stable-rank activations changes whose principal directions approximate the refusal subspace.

B Limitations and assumptions of analytical derivations

Throughout the derivations, we make simplifying assumptions to obtain a tractable model of the relationship between refusal targets, gradients, and activation changes.

Supervised refusal-training scope. Our derivation models supervised fine-tuning on harmful prompts paired with textual refusals. We use the terms refusal training and supervised safety tuning for this setting. DPO, RLHF, and RLVR use different objectives and are not directly covered by the derivation which may induce slightly. The controlled fine-tuning experiment uses effective batch size one to isolate per-example refusal updates. It is intended as an approximation to the sporadic contribution of comparatively rare harmful-refusal examples within SFT dataset mixture for our target model. Larger batches containing unrelated instruction-following examples may introduce additional signal into gradients and the resulting activation changes.

First-token-only loss. We compute all losses only at the first assistant target-token position, after chat formatting, with the prompt positions masked. We therefore omit gradients from the remaining completion tokens. In the evaluated settings, the first-token loss captures substantial refusal-direction structure, but later target positions may introduce additional directions and training effects.

Deterministic per-example updates. We compute isolated per-example gradient steps with deterministic backward passes and without stochastic regularization such as dropout. These measurements do not reproduce minibatch interactions, optimizer momentum, adaptive preconditioning, weight decay, or the accumulation of many training steps.

Prompt-dependent Jacobians and conditioning. Exact transformer Jacobians are prompt-dependent, including through input-dependent attention patterns. Consequently, Theorem 3.2 is an idealized sufficient condition rather than a claim that complete per-example Jacobians are shared. Products of worst-case condition-number bounds may also become numerically vacuous for anisotropic maps. The reduced-map condition numbers reported in our experiments apply only to selected data-dependent source coordinates and are not estimates of the full prompt-specific Jacobian condition numbers.

In-sample shared-map fit and predictive scope. When $\epsilon_{\max}$ is obtained by measuring the transfer-factor fit error on the same N examples used to estimate $\hat{J}_k^{L \rightarrow l}$, Lemma 3.6 gives an in-sample fit-conditioned interval for $\text{sr}(G^{(l)})$, rather than an independent prediction. Specifically, it certifies that, on the fitted support, the map-implied transfer factor $\hat{\tau}_k^{L \rightarrow l}$ agrees with the observed transfer factor $\tau^{L \rightarrow l}$ to relative error at most $\epsilon_{\max}$, and therefore that the observed target stable rank lies within the corresponding interval. The result becomes predictive only when the same fitted map is held fixed and applied without refitting to additional source gradient matrices, with its error bound established independently on held-out source-target matrix pairs. Moreover, agreement at the level of stable rank does not imply matrix-level agreement; we evaluate the latter separately using relative Frobenius error, relative spectral error, and row-wise cosine similarity.

C Additional details about experimental setup.

Refusal-mediating layers. Unless otherwise stated, the frozen-model single-layer rank-vulnerability and ablation analyses use layer 7 since our measurements show that refusal vector estimation at that layer provides most effective, and therefore attacker-preferred refusal ablation. Layerwise profiles and cross-layer transfer experiments additionally report all explicitly labeled layers. The controlled fine-tuning rank-vulnerability experiment in Section 4.6 uses layer 8 – the earliest layer in the range of effective layers in our experiment, which retains a long last-to-middle-layer path over which to test rank propagation, similar to frozen-model experiments; selecting a later layer would make that propagation test less stringent.

Estimating gradient-induced activation updates. To compute gradient-induced activation updates $\mathcal{G}$, for each training example i and cross-entropy loss ℓ_i^{CE} on the first token of the target completion, we apply one small parameter update

$$\theta_{\text{new}} = \theta - \eta \nabla_{\theta} \ell_i^{\text{CE}}, \quad \eta = 10^{-6}.$$

At layer l , we measure

$$\mathbf{g}_i^{(l)} := - \left(h_i^{(l)}(\theta_{\text{new}}) - h_i^{(l)}(\theta) \right),$$

and then restore the parameters to θ . Thus, $\mathbf{g}_i^{(l)}$ is the negative realized activation change from one small gradient step, consistent with our sign convention.

For the evaluated models, we use $l = 7$ or $l = 8$, the earliest layers at which the estimated refusal directions yield substantial attack success; this retains a long last-to-middle-layer path over which to test rank propagation. All empirical $\mathcal{G}$ matrices are computed from the realized finite differences in Equation (15).

Under a local first-order expansion, their corresponding differential is

$$\eta J_i^{(l)} \nabla_{\theta} \ell_i^{\text{CE}},$$

which can instead be evaluated directly as a Jacobian–vector product without applying a parameter perturbation. We leave a systematic comparison of the difference and differential estimators to future work.

D Broader limitations

Model, data, language, and checkpoint scope. This work is primarily a case study of one small OLMo model family, relatively small prompt sets, and English-language data. We do not identify an attackable difference-in-means refusal direction in the base checkpoint under our estimator, but this does not rule out other functional base-model refusal mechanisms or representations. The released Base, SFT, DPO, RLVR1, and Instruct checkpoints differ in their objectives, data, optimization histories, and numbers of updates. Their comparison is therefore descriptive and does not identify which training stage caused the observed refusal geometry or change in ablation sensitivity.

Target-set-adapted white-box threat model. For each evaluated prompt set S , the attacker estimates one difference-in-means direction from all prompts in S and uses that direction for the refusal ablation attack on S . The results therefore intentionally test an attacker-favorable same-set ablation setting and do not establish that a direction estimated from one prompt sample transfers to unseen prompts. Held-out universality of refusal directions is a distinct question left for future work.

Attack-family scope. Our primary attack is difference-in-means single-vector ablation. SAE-based steering, multi-feature ablation, and broader automated representation-level attacks remain outside our evaluation. We approach the difference-in-means single-vector attack using analytical methods and invite application of those methods to study a broader set of attacks and defenses in mechanistic interpretability.

Stable rank versus causal mediation and intrinsic dimension. Stable rank measures global linear spectral concentration. It does not by itself measure refusal-signal magnitude, causal mediation strength, or local nonlinear manifold dimension. A layer may therefore have relatively high stable rank while its refusal direction has little behavioral effect, whereas another layer may have lower stable rank but mediate refusal strongly. Our rank–vulnerability comparisons concern fixed or matched refusal-mediating layers and do not use stable rank to predict which layer is most causally important.

Scope beyond refusal. We study refusal because it provides a shared refuse/comply mechanism with a well-defined white-box ablation attack. It remains unclear whether similarly universal non-refusal concepts have the same geometry, or whether increasing feature rank generally makes universal steering or other forms of jailbreak optimization harder.

E Theorems

E.1 Stable rank of the refusal first-token matrix

Definition E.1 (One-hot matrix E_r for a batch of refusals). Let $E_r \in \mathbb{R}^{N \times |V|}$ be the row-stacked one-hot matrix with i -th row $e_{r_i}^\top$, for a batch of targets $r = (r_1, \dots, r_N)$. Let $c_k := \#\{i : r_i = k\}$ be the count of token k , let $f_k := c_k/N$, and let $m := |\{k : c_k > 0\}|$.

Lemma E.2 (Spectrum of a stacked one-hot matrix). *The feature-space Gram matrix of E_r is diagonal:*

$$E_r^\top E_r = \text{diag}(c_1, \dots, c_{|V|}). \quad (33)$$

Consequently, the nonzero singular values of E_r are $\sqrt{c_k}$ for tokens with $c_k > 0$, and

$$\text{rank}(E_r) = m. \quad (34)$$

The stable rank is

$$\text{sr}(E_r) = \frac{\|E_r\|_F^2}{\|E_r\|_2^2} = \frac{\sum_{k=1}^{|V|} c_k}{\max_k c_k} = \frac{N}{\max_k c_k} = \frac{1}{\max_k f_k}. \quad (35)$$

Proof. The proof uses an approach similar to [44]. For any $k, k' \in V$,

$$(E_r^\top E_r)_{k,k'} = \sum_{i=1}^N (E_r)_{i,k} (E_r)_{i,k'} = \sum_{i=1}^N \mathbf{1}[r_i = k] \mathbf{1}[r_i = k']. \quad (36)$$

If $k \neq k'$, no example can satisfy both $r_i = k$ and $r_i = k'$, so the sum is 0. If $k = k'$, the sum is c_k . Thus $E_r^\top E_r = \text{diag}(c_1, \dots, c_{|V|})$. The eigenvalues of $E_r^\top E_r$ are c_k , so the singular values of E_r are $\sqrt{c_k}$. The rank is the number of positive counts. Finally, $\|E_r\|_F^2 = \sum_k c_k = N$, while $\|E_r\|_2^2 = \max_k c_k$, giving the stable-rank formula. $\square$

Corollary E.3 (Minimizer and maximizer of $\text{sr}(E_r)$). *Let $c_{\max} := \max_k c_k$. Then*

$$1 \leq \text{sr}(E_r) = \frac{N}{c_{\max}} \leq m. \quad (37)$$

Moreover:

1. $\text{sr}(E_r) = 1$ iff $c_{\max} = N$, i.e., all targets share the same first token.
2. For fixed support size m , $\text{sr}(E_r)$ is maximized when counts are as balanced as possible across the m observed tokens, i.e. $c_k \in \{\lfloor N/m \rfloor, \lceil N/m \rceil\}$ on the support. In that case,

$$\max \text{sr}(E_r) = \frac{N}{\lceil N/m \rceil}. \quad (38)$$

If $m \mid N$, this maximum equals m .

3. If the support is unconstrained and $N \leq |V|$, the global maximum is $\text{sr}(E_r) = N$, achieved when all first tokens are distinct. If $N > |V|$, the maximum is achieved by balancing counts over all $|V|$ vocabulary tokens.

Proof. By Lemma E.2, $\text{sr}(E_r) = N/c_{\max}$. The lower bound follows from $c_{\max} \leq N$, with equality iff $c_{\max} = N$. For fixed support size m , the pigeonhole principle gives $c_{\max} \geq \lceil N/m \rceil$, hence

$$\text{sr}(E_r) = \frac{N}{c_{\max}} \leq \frac{N}{\lceil N/m \rceil} \leq m. \quad (39)$$

This is tight when counts are balanced across the m observed targets. The unconstrained-support statement follows by taking $m = \min\{N, |V|\}$. $\square$

E.2 Stable rank propagation theorems

The following proof uses standard matrix norm inequalities: the Frobenius/Schatten-2 ideal property, operator norm submultiplicativity, and the definition of the spectral condition number; see, e.g., [1, Chapter 5].

Theorem E.4 (Stable rank under multiplication by a well-conditioned factor). *Let $A \in \mathbb{R}^{N \times n}$, let $B \in \mathbb{R}^{n \times n}$ be invertible, and define condition number:*

$$\kappa(B) := \|B\|_2 \|B^{-1}\|_2. \quad (40)$$

Then

$$\frac{1}{\kappa(B)^2} \text{sr}(A) \leq \text{sr}(AB) \leq \kappa(B)^2 \text{sr}(A). \quad (41)$$

Proof. Recall that

$$\text{sr}(M) := \frac{\|M\|_F^2}{\|M\|_2^2}. \quad (42)$$

We use

$$\|AB\|_F \leq \|A\|_F \|B\|_2, \quad (43)$$

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2. \quad (44)$$

Since $A = (AB)B^{-1}$,

$$\|A\|_2 = \|(AB)B^{-1}\|_2 \leq \|AB\|_2 \|B^{-1}\|_2, \quad (45)$$

so

$$\|AB\|_2 \geq \frac{\|A\|_2}{\|B^{-1}\|_2}. \quad (46)$$

Therefore,

$$\text{sr}(AB) = \frac{\|AB\|_F^2}{\|AB\|_2^2} \leq \frac{(\|A\|_F \|B\|_2)^2}{(\|A\|_2 / \|B^{-1}\|_2)^2} = \kappa(B)^2 \text{sr}(A). \quad (47)$$

For the lower bound, apply the upper bound to $A = (AB)B^{-1}$:

$$\text{sr}(A) \leq \kappa(B^{-1})^2 \text{sr}(AB) = \kappa(B)^2 \text{sr}(AB). \quad (48)$$

Rearranging gives

$$\text{sr}(AB) \geq \frac{1}{\kappa(B)^2} \text{sr}(A). \quad (49)$$

□

Lemma E.5 (Stable rank bounds from transfer factor fit error). *Let $G^{(L)}, G^{(l)} \in \mathbb{R}^{N \times d}$ be nonzero gradient matrices, and let $\hat{J}_k^{L \rightarrow l}$ be a top- k source-subspace shared map with fitted output*

$$\hat{G}_k^{(l)} = G^{(L)} \hat{J}_k^{L \rightarrow l}.$$

Suppose that the relative transfer-factor fit error is bounded by some ϵ_{max} :

$$\epsilon_{\tau,k}^{L \rightarrow l} \leq \epsilon_{max} < 1.$$

Then

$$\frac{\hat{\tau}_k^{L \rightarrow l}}{1 + \epsilon_{max}} \text{sr}(G^{(L)}) \leq \text{sr}(G^{(l)}) \leq \frac{\hat{\tau}_k^{L \rightarrow l}}{1 - \epsilon_{max}} \text{sr}(G^{(L)}).$$

Proof. By the definition of the relative transfer-factor fit error and assuming it is bounded by ϵ_{max} ,

$$|\hat{\tau}_k^{L \rightarrow l} - \tau^{L \rightarrow l}| \leq \epsilon_{max} \tau^{L \rightarrow l}.$$

Hence,

$$(1 - \epsilon_{max}) \tau^{L \rightarrow l} \leq \hat{\tau}_k^{L \rightarrow l} \leq (1 + \epsilon_{max}) \tau^{L \rightarrow l}.$$

Since $\epsilon_{max} < 1$, rearranging gives

$$\frac{\hat{\tau}_k^{L \rightarrow l}}{1 + \epsilon_{max}} \leq \tau^{L \rightarrow l} \leq \frac{\hat{\tau}_k^{L \rightarrow l}}{1 - \epsilon_{max}}.$$

Finally, using

$$\text{sr}(G^{(l)}) = \tau^{L \rightarrow l} \text{sr}(G^{(L)})$$

proves the result. □

F Matrix metrics

This appendix defines the matrix metrics used in the empirical sections. All matrices below are row-stacked matrices in $\mathbb{R}^{n \times d}$, such as benign-centered refusal residuals ΔH or gradient-induced activation update matrices $\mathcal{G}^{(l)}$.

Thin SVD and right-singular subspaces. For $M \in \mathbb{R}^{n \times d}$, let

$$M = U_M \Sigma_M V_M^\top$$

denote its thin SVD, where $V_M \in \mathbb{R}^{d \times r_M}$ contains the right singular vectors associated with nonzero singular values and $r_M := \text{rank}(M)$. We write

$$\mathcal{V}(M) := \text{span}(V_M)$$

Uncentered single-matrix metrics. For metrics of a single matrix, we use the uncentered matrix M . The empirical second-moment matrix is

$$C_M := \frac{1}{n} M^\top M, \quad \lambda_{M,i} := \frac{\sigma_{M,i}^2}{n}.$$

The stable rank is

$$\text{sr}(M) := \frac{\|M\|_F^2}{\|M\|_2^2} = \frac{\sum_{i=1}^{r_M} \sigma_{M,i}^2}{\sigma_{M,1}^2}.$$

Centered pairwise comparisons between right singular subspaces. When comparing right singular subspaces between matrices, we use centered matrices, which means that for each matrix M , we remove its row mean μ_M to obtain its centered matrix M_c before computing its right singular vectors:

$$\mu_M = \frac{1}{n} M^\top \mathbf{1}_n, \quad M_c = M - \mathbf{1}_n \mu_M^\top.$$

Unless otherwise stated, $V_{M,k}$ and $\lambda_{M,i}$ in pairwise metrics are computed from the centered matrix M_c . This separates the mean direction, measured by cosine similarity, from the remaining PCA-style subspace structure.

Top- k policy. For two matrices A, B , let $r_A := \text{rank}(A_c)$ and $r_B := \text{rank}(B_c)$. Given a requested k_{top} , we use

$$k := \min\{k_{\text{top}}, r_A, r_B\}.$$

In the main experiments we set $k_{\text{top}} = 100$. Since $n < 100$ for our experiments, this includes all nonzero centered principal directions after the rank cap.

Cosine similarity of row means. For $A, B \in \mathbb{R}^{n \times d}$, define

$$\text{cosine_mean}(A, B) := \frac{\mu_A^\top \mu_B}{\|\mu_A\|_2 \|\mu_B\|_2} \in [-1, 1].$$

This metric compares only the mean directions and is complementary to centered subspace-overlap metrics.

Explained-variance overlap. Let $A_c = U_A \Sigma_A V_A^\top$ and $B_c = U_B \Sigma_B V_B^\top$ be centered thin SVDs. Let $V_{A,k}$ and $V_{B,k}$ denote the first k right singular vectors, and define

$$C_{A,B}^{(k)} := V_{A,k}^\top V_{B,k}.$$

The explained-variance overlap of A 's top- k directions captured by B 's top- k subspace is

$$\text{EV}_k(A \mid B) := \sum_{i=1}^k \frac{\lambda_{A,i}}{\sum_{\ell=1}^k \lambda_{A,\ell}} \sum_{j=1}^k \left(C_{A,B}^{(k)} \right)_{ij}^2 \in [0, 1].$$

Equivalently, each A -principal direction is weighted by its share of A 's top- k variance and scored by the squared norm of its projection onto $\text{span}(V_{B,k})$. The metric is directional because it weights only by A 's spectrum.

Average principal-angle overlap. Let

$$A_c = U_A \Sigma_A V_A^\top, \quad B_c = U_B \Sigma_B V_B^\top$$

be centered thin SVDs. Let $V_{A,k}$ and $V_{B,k}$ contain the first k right singular vectors, and define

$$C_{A,B}^{(k)} := V_{A,k}^\top V_{B,k}.$$

Let $s_i(C_{A,B}^{(k)})$ denote the singular values of $C_{A,B}^{(k)}$. We define the average principal-angle overlap as

$$\text{AvgOverlap}_k(A, B) := \frac{1}{k} \sum_{i=1}^k s_i(C_{A,B}^{(k)})^2 \quad (50)$$

$$= \frac{1}{k} \|V_{A,k}^\top V_{B,k}\|_F^2 \in [0, 1]. \quad (51)$$

The singular values of $C_{A,B}^{(k)}$ are the cosines of the principal angles between the two subspaces. Thus, AvgOverlap_k is symmetric and basis-invariant, equals 1 when the top- k subspaces coincide, and equals 0 when they are orthogonal.

Shared-map metrics. All shared-map metrics below are computed on uncentered gradient matrices. Let $G^{(L)}, G^{(l)} \in \mathbb{R}^{N \times d}$ be the source and target gradient matrices, and let

$$\hat{G}_k^{(l)} = G^{(L)} \hat{J}_k^{L \rightarrow l}$$

be the output of the fitted top- k source-subspace shared map.

Source-subspace energy capture. Let $V_k^{(L)} \in \mathbb{R}^{d \times k}$ contain the top- k right singular vectors of $G^{(L)}$. We define the fraction of source-gradient energy contained in the selected subspace as

$$\text{EC}_k^{(L)} := \frac{\|G^{(L)} V_k^{(L)}\|_F^2}{\|G^{(L)}\|_F^2}. \quad (52)$$

Equivalently, if $\sigma_1(G^{(L)}) \geq \sigma_2(G^{(L)}) \geq \dots$ are the singular values of $G^{(L)}$, then

$$\text{EC}_k^{(L)} = \frac{\sum_{j=1}^k \sigma_j(G^{(L)})^2}{\sum_j \sigma_j(G^{(L)})^2}. \quad (53)$$

This metric measures how much of the total squared source-gradient energy is retained before fitting the shared map.

Relative Frobenius reconstruction error. We define the relative Frobenius error between the fitted and observed target matrices as

$$\text{RelF}_k^{L \rightarrow l} := \frac{\|G^{(l)} - \hat{G}_k^{(l)}\|_F}{\|G^{(l)}\|_F}. \quad (54)$$

This metric measures the total reconstruction error relative to the Frobenius norm of the observed target matrix. A value of zero denotes exact reconstruction; the metric is not upper bounded by one.

Relative spectral reconstruction error. We define the relative spectral error as

$$\text{Rel2}_k^{L \rightarrow l} := \frac{\|G^{(l)} - \hat{G}_k^{(l)}\|_2}{\|G^{(l)}\|_2}. \quad (55)$$

This metric measures the largest residual singular direction relative to the dominant singular direction of the observed target matrix. It complements the relative Frobenius error because stable rank depends on both the Frobenius and spectral norms.

F.1 Refusal direction across released post-training checkpoints - additional plots

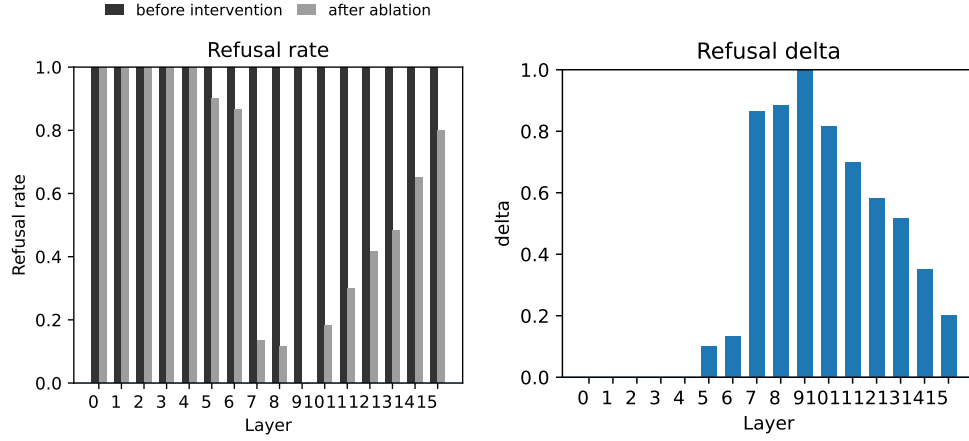


Figure 12: Refusal ablation attack effectiveness per each layer used to estimate the refusal difference-in-means direction (for a fixed model OLMo-2-0425-1B-Instruct against a subset of AdvBench prompts). The plot shows that fixing layers to 7-10 provides the largest ablation attack success (refusal delta) with decreasing success across later layers, while earlier layers provide almost no benefit, resulting in relatively ineffective attacks.

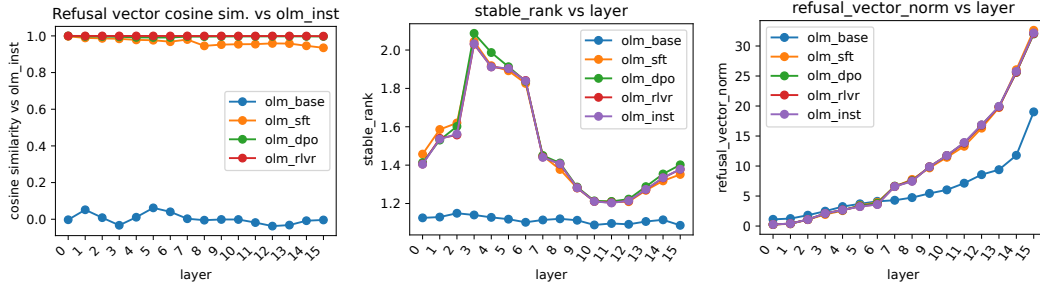


Figure 13: Measuring refusal vectors across stages shows that once a functional refusal vector appears at SFT stage, all resulting refusal directions across all stages (except the base model) are all similar for a subset of selected AdvBench prompts. Refusal vector norms grow after SFT stages and stay similar throughout post-training stages.

G Refusal first-token diversity increases gradient stable ranks - additional plots

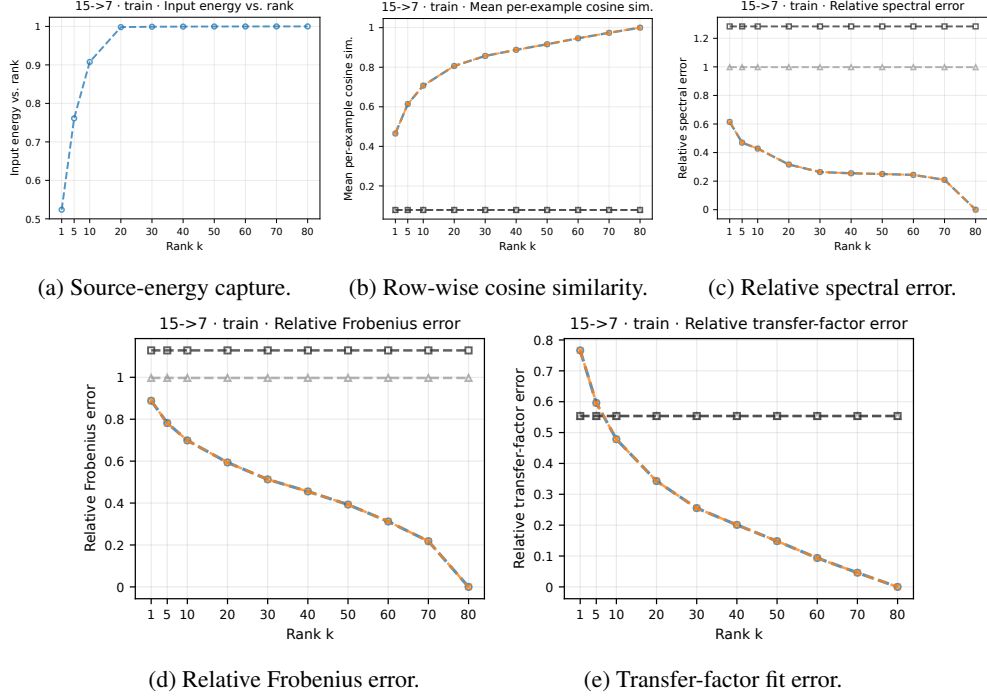


Figure 14: In-sample fit quality of the top- k source-subspace shared map from $G^{(15)}$ to $G^{(7)}$ for the $m = 16$ refusal-start condition. Increasing k captures more source-gradient energy, improves per-example directional agreement, and reduces matrix-reconstruction and stable-rank-transfer errors. At $k = 20$, the map captures 99.8% of source energy and attains relative Frobenius, spectral, and transfer-factor errors 0.594, 0.316, and 0.343. The $k = 80$ endpoint is the in-sample interpolation limit and is shown only as a reference, not as evidence of a low-rank mechanism.

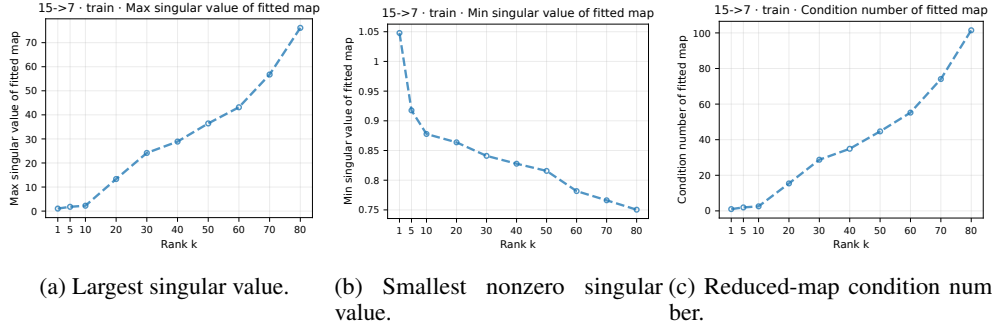


Figure 15: Singular-value diagnostics for the fitted map on its selected source coordinates. The smallest nonzero singular value changes little between $k = 10$ and $k = 20$, from 0.878 to 0.864, while the largest singular value grows from 2.24 to 13.33. Consequently, the reduced-map condition number rises from 2.55 to 15.43. This explains why the higher-rank map gives a better empirical fit while its worst-case condition-number bound becomes less informative.

H Refusal first-token diversity increases residual rank and weakens single-vector ablation - additional plots

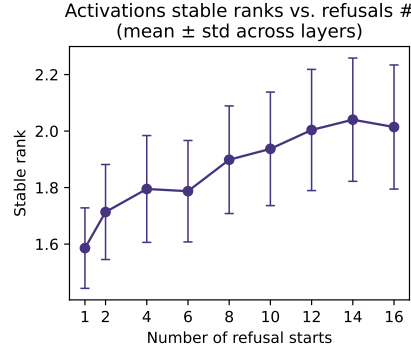


Figure 16: Averaged ranks over layers 7-15 stable rank of benign-centered refusal residuals ΔH . Increasing refusal first-token diversity raises $\text{sr}(\Delta H)$ in middle-to-late layers.

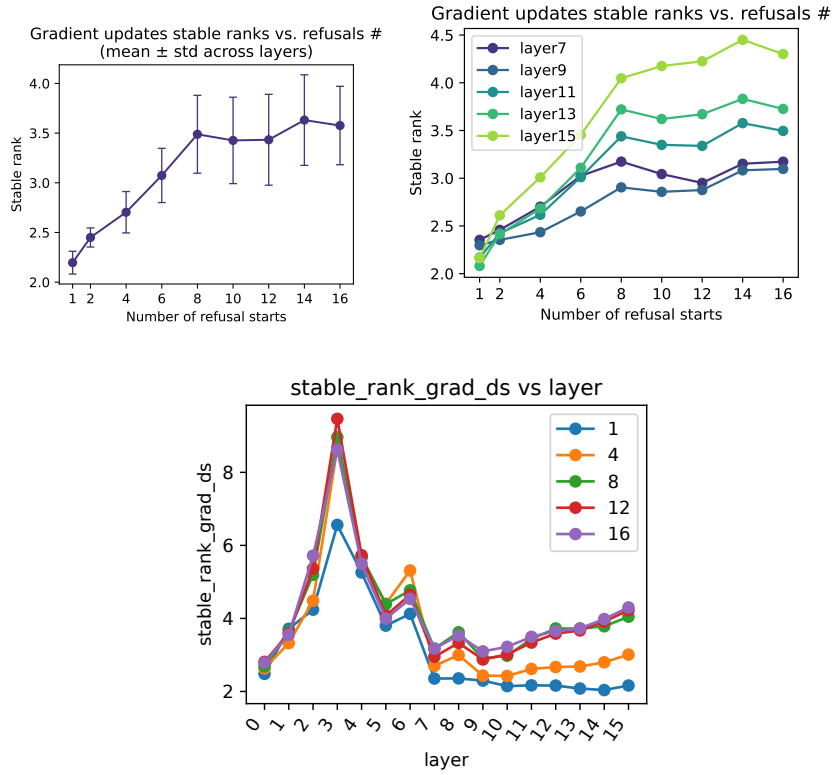


Figure 17: Stable rank of gradient-induced update deltas $\mathcal{G}$: more refusal first-token buckets produce higher effective rank across most layers, although the increase saturates at the end. Left upper plot averages stable ranks across layers 7-15.

H.1 Frozen-model raw-gradient shared-map diagnostics

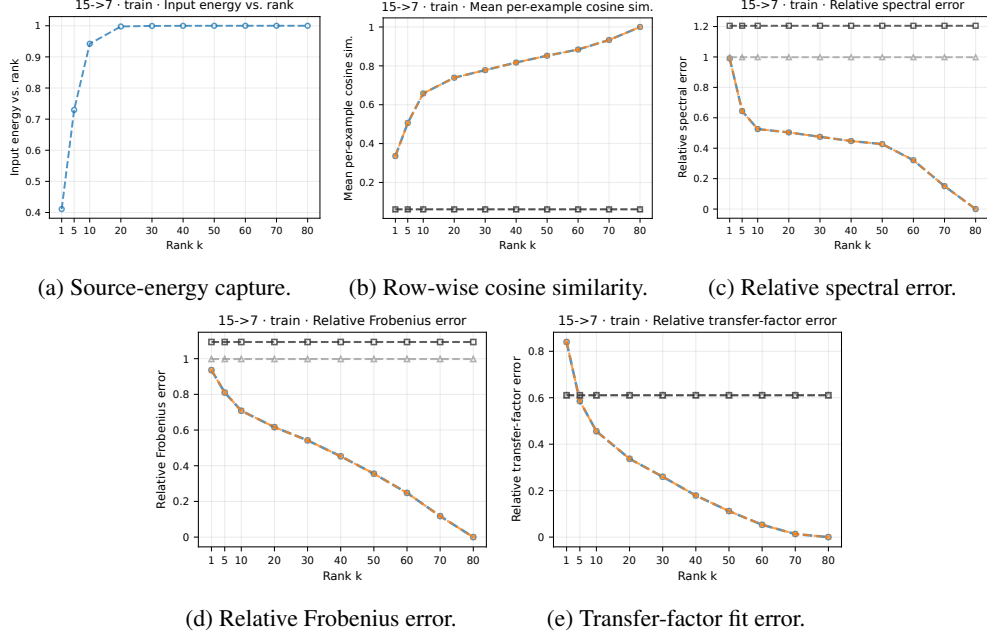


Figure 18: In-sample fit quality of the top- k source-subspace shared map from $G^{(15)}$ to $G^{(7)}$ for the $m = 16$ prompt-subset condition. Increasing k captures more source-gradient energy, improves per-example directional agreement, and reduces matrix-reconstruction and stable-rank-transfer errors. At $k = 20$, the selected source subspace captures 99.8% of source-gradient energy. The relative Frobenius and spectral errors are 0.616 and 0.504, the mean row-wise cosine similarity is 0.739, and the relative transfer-factor fit error is 0.337. Thus, the rank-20 source subspace contains nearly all source-gradient energy, while the remaining reconstruction errors show that the shared map captures a meaningful but incomplete component of the observed cross-layer relation. The $k = 80$ endpoint is the in-sample interpolation limit and is shown only as a reference, not as evidence of a low-rank mechanism.

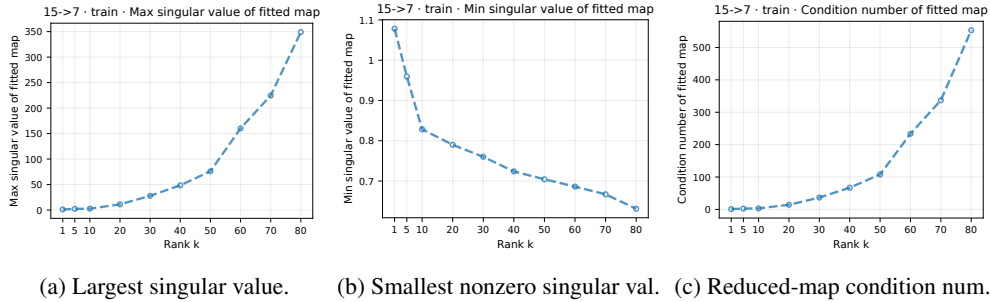


Figure 19: Singular-value diagnostics for the reduced coordinate map fitted to the raw gradients. Between $k = 10$ and $k = 20$, the smallest nonzero singular value decreases slightly from 0.83 to 0.79, the largest singular value grows from 2.65 to 11.19, the reduced-map condition number rises from 3.19 to 14.17. The higher-rank map gives a better in-sample fit but becomes substantially more anisotropic, making its worst-case condition-number bound less informative.

H.2 Frozen-model gradient-update shared-map diagnostics

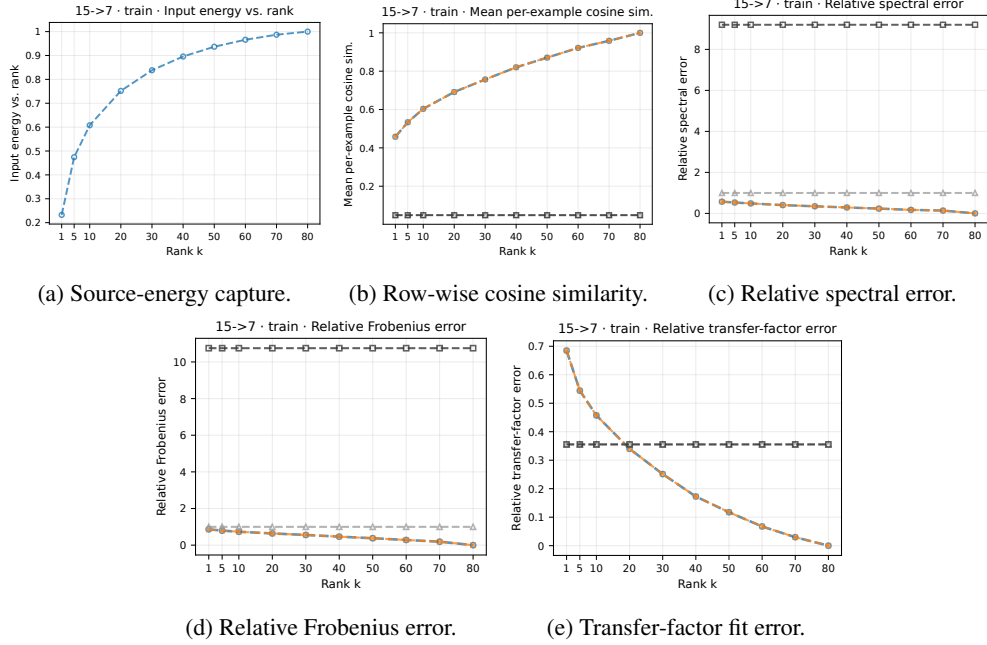


Figure 20: In-sample fit quality of the top- k source-subspace shared map from $\mathcal{G}^{(15)}$ to $\mathcal{G}^{(7)}$ for the $m = 16$ prompt-subset condition. Increasing k captures more source-update energy, improves per-example directional agreement, and reduces matrix-reconstruction and stable-rank-transfer errors. At $k = 20$, the selected source subspace captures 75.2% of source-update energy. The relative Frobenius and spectral errors are 0.639 and 0.405, the mean row-wise cosine similarity is 0.691, and the relative transfer-factor fit error is 0.340. The more gradual energy-capture curve indicates that substantial source-update energy lies beyond the first 20 principal directions. The $k = 80$ endpoint is the in-sample interpolation limit and is shown only as a reference, not as evidence of a low-rank mechanism.

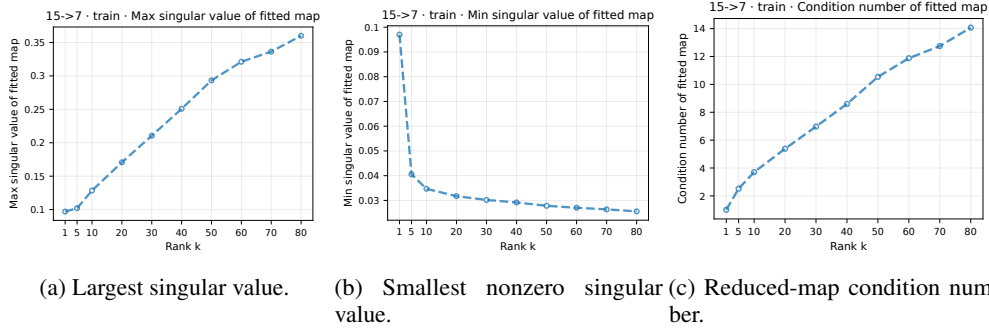


Figure 21: Singular-value diagnostics for the reduced coordinate map fitted to the gradient-induced activation updates. Between $k = 10$ and $k = 20$, the smallest nonzero singular value decreases from 0.0347 to 0.0317, while the largest singular value grows from 0.129 to 0.171. Consequently, the reduced-map condition number rises from 3.71 to 5.38. The rank-20 map improves the in-sample fit with substantially less conditioning deterioration than the corresponding raw-gradient map.

I Diverse-refusal fine-tuning raises refusal residuals rank and weakens ablation - additional plots

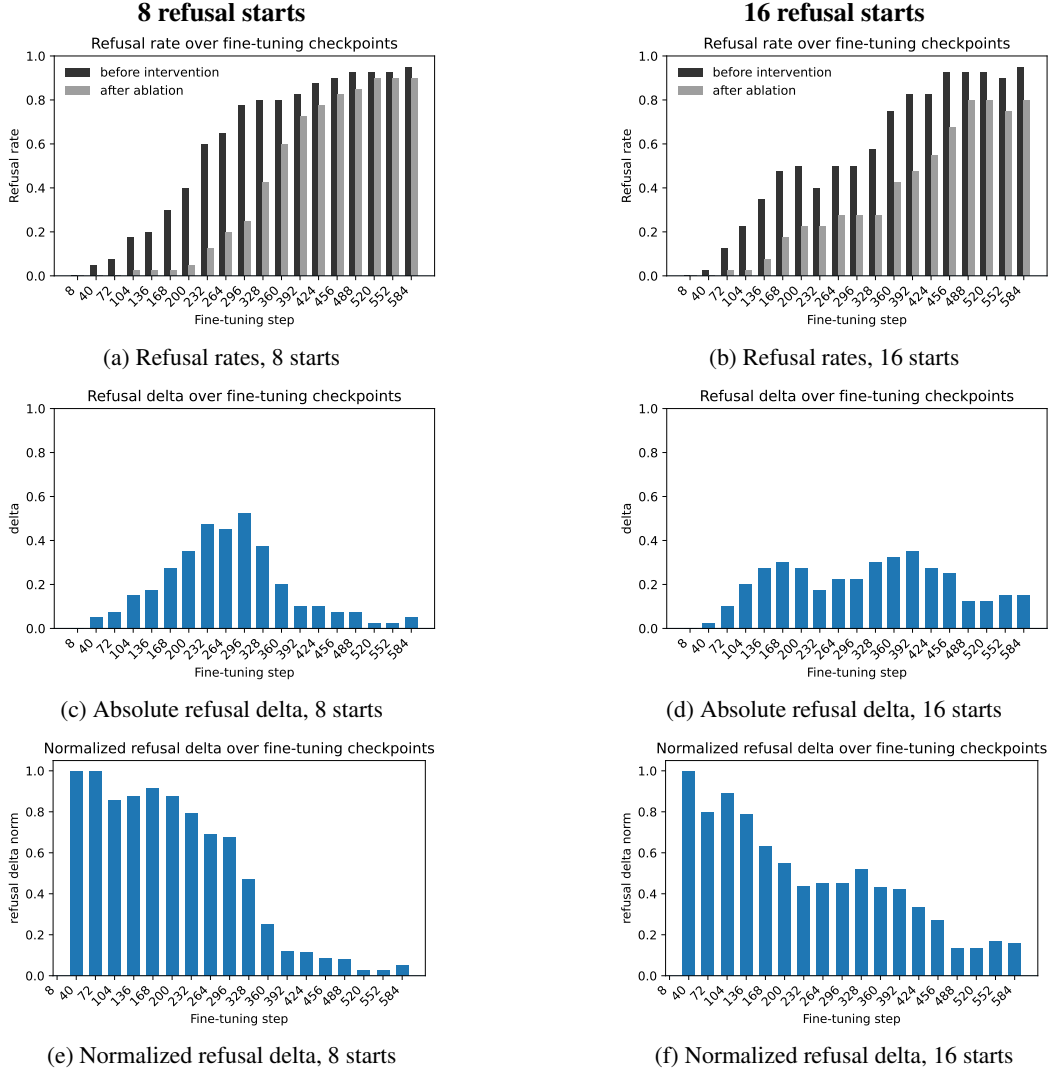


Figure 22: Optimization-step-matched refusal ablation trajectories for models trained with 8 and 16 refusal starts. Left and right columns show the 8- and 16-start conditions, respectively. The top row reports baseline and post-ablation refusal rates; the middle row reports the absolute refusal deltas; and the bottom row reports the fraction of baseline refusals removed by ablation. Ablation vulnerability exists in a small window of training since the refusal delta is peaked at first when maximum refusal rate is reached, decaying afterwards. The 8-start plot shows deltas having a single peak with maximum of 0.55, while 16-start plot shows two smaller peaks and reaches overall maximum of 0.375, consistent with reduced ablation attack effectiveness.

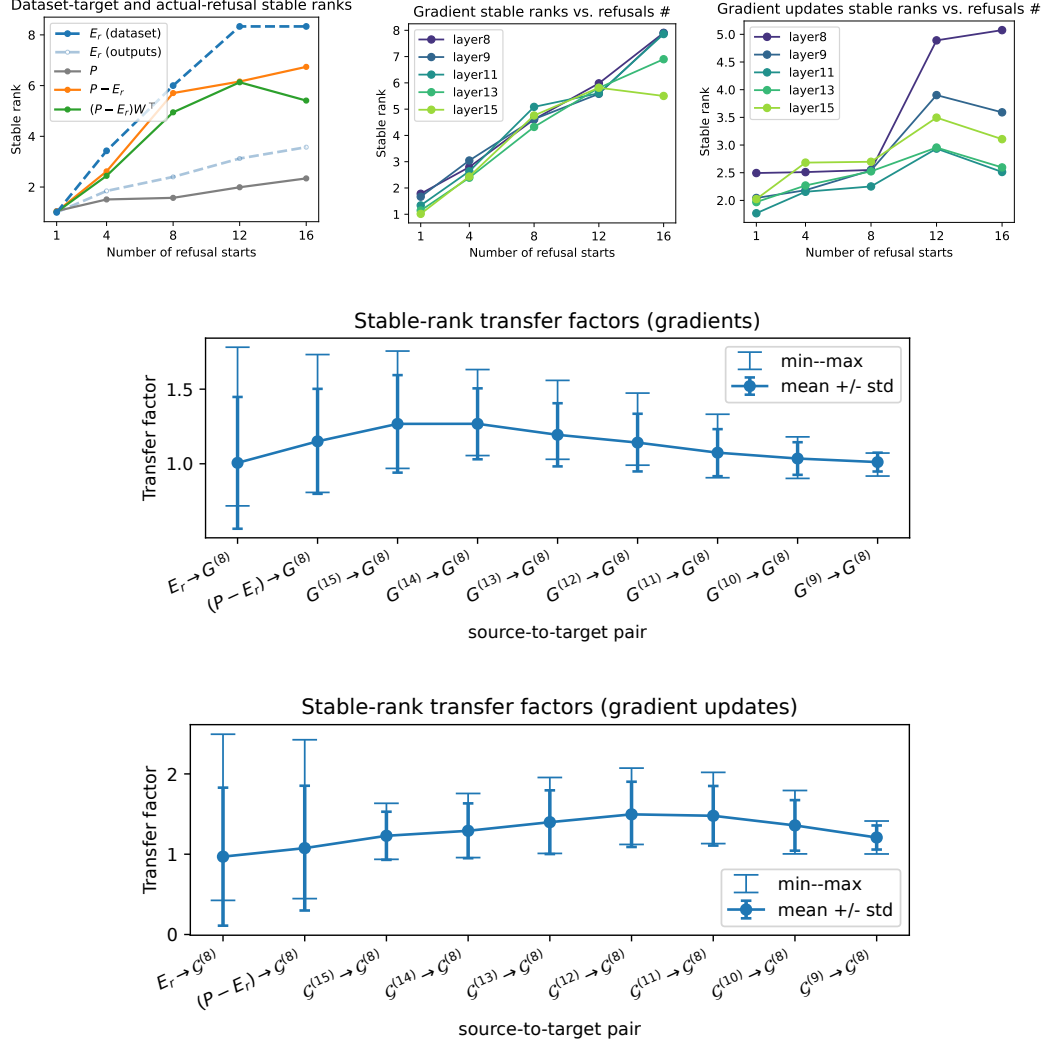


Figure 23: Stable ranks and stable rank propagation under refusal-start diversification by fine-tuning for gradient and gradient-induced activation deltas. Top left: stable ranks of the dataset-target E_r , realized-output E_r , prediction matrix P , target residual $P - E_r$, and resulting last-layer gradient $(P - E_r)W^\top$. Top middle: stable ranks of the gradients $G^{(l)}$. Top right: stable ranks of the gradient-induced activation updates $\mathcal{G}^{(l)}$. Bottom rows: end-to-end stable rank transfer factors measured for raw gradients and gradient-induced activation deltas transfer factors. The target-space, gradients, and gradient-updates stable ranks increase overall as refusal-start support broadens, and the both gradient and gradient-update matrices undergo mild stable-rank expansion from layer 15 to layer 8 on average.