SOM Directions Are Better than One: Multi-Directional Refusal Suppression in Language Models

Giorgio Piras[*1], **Raffaele Mura**[*1], **Fabio Brau**[1], **Luca Oneto**[2], **Fabio Roli**[2], **Battista Biggio**[1]
[1]University of Cagliari
[2]University of Genova
{giorgio.piras, raffaele.mura, fabio.brau, battista.biggio}@unica.it, {luca.oneto, fabio.roli}@unige.it

Abstract

Refusal refers to the functional behavior enabling safety-aligned language models to reject harmful or unethical prompts. Following the growing scientific interest in mechanistic interpretability, recent work encoded refusal behavior as a single direction in the model's latent space; e.g., computed as the difference between the centroids of harmful and harmless prompt representations. However, emerging evidence suggests that concepts in LLMs often appear to be encoded as a low-dimensional manifold embedded in the high-dimensional latent space. Motivated by these findings, we propose a novel method leveraging Self-Organizing Maps (SOMs) to extract multiple refusal directions. To this end, we first prove that SOMs generalize the prior work's difference-in-means technique. We then train SOMs on harmful prompt representations to identify multiple neurons. By subtracting the centroid of harmless representations from each neuron, we derive a set of multiple directions expressing the refusal concept. We validate our method on an extensive experimental setup, demonstrating that ablating multiple directions from models' internals outperforms not only the single-direction baseline but also specialized jailbreak algorithms, leading to an effective suppression of refusal. Finally, we conclude by analyzing the mechanistic implications of our approach.

Code — https://github.com/pralab/som-refusal-directions
Extended version — https://arxiv.org/pdf/2511.08379

1 Introduction

The use of extensive data while training Large Language Models (LLMs) introduces safety challenges, as harmful content is inevitably included and exposes models to potential misuse through the elicitation of restricted outputs (Wei et al. 2022; Carlini et al. 2023). While safety alignment procedures are designed to mitigate these risks and enable models to refuse harmful prompts (Bai et al. 2022; Touvron et al. 2023), LLMs remain susceptible to jailbreak attacks bypassing such safeguards (Zou et al. 2023; Andriushchenko, Croce, and Flammarion 2025). In response, a growing body of work has sought to explore why safety alignment fails by connecting such a *refusal behavior* to models' activation space. Following the linear representation hypothesis (Park, Choe, and Veitch 2024), positing that concepts are represented as single linear directions, refusal has been encoded by subtracting the centroid of harmless prompt representations from that of harmful ones. This single direction (SD), capturing the shift in model representations associated with refusal, has been shown to induce jailbreak when ablated from the model's internals (Arditi et al. 2024). However, advances in mechanistic interpretability have questioned the general adequacy of the single-direction view, showing that semantic and functional concepts—such as days (Engels et al. 2025), trigonometric (Kantamneni and Tegmark 2025), and other broader concepts (Modell, Rubin-Delanchy, and Whiteley 2025)—are not captured by a single direction, but instead span low-dimensional manifolds in high-dimensional spaces. These manifolds are often composed of multiple, closely related directions that jointly express different facets of a concept, challenging the single-direction view and motivating new methodologies to uncover and characterize these multi-dimensional perspectives (Olah and Jermyn 2024). Despite this paradigm shift, existing work on refusal is mostly confined to the single-direction framework (Arditi et al. 2024), or focused on orthogonal multiple directions that may represent disjoint subcomponents of refusal (Wollschläger et al. 2025; Pan et al.

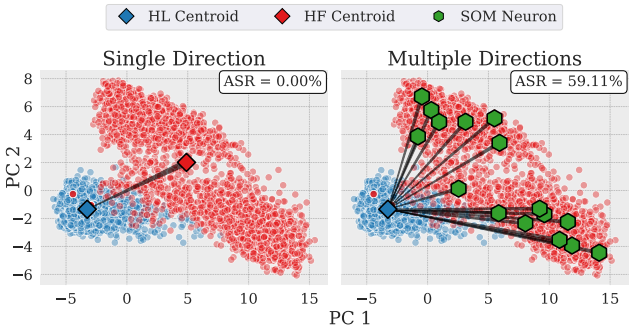


Figure 1: Single and multiple directions in the representation space of Llama2-7B. While SD (left) captures a single view of refusal, our MD (right) approach, via a 4x4 SOM, enables a multi-faceted perspective of refusal, and, thus, a higher Attack Success Rate (ASR).

[*]These authors contributed equally.

2025). These approaches overlook that refusal may be encoded into a higher-dimensional manifold spanned by multiple directions in representation space.

To address this gap, we propose a novel multi-directional approach for encoding and suppressing refusal behavior in LLMs. Central to our method are Self-Organizing Maps (SOMs), whose ability to capture high-dimensional structures enables the discovery of multiple closely related directions, facilitating more effective refusal suppression. In detail, following the SD design, based on the difference between harmless and harmful prompts' centroids, we first prove how a single neuron SOM generalizes the standard centroid computation. Then, we present our multi-directional (MD) approach, previewed in Fig. 1. Specifically, we train a SOM on internal representations collected at the generation step following harmful prompts, where refusal behavior is first expressed in the model's activations. Then, each SOM neuron encoding a localized region of the harmful distribution is translated into a direction by subtracting the harmless centroid, thus generalizing the SD measure, and obtaining a set of candidate directions for ablation. We then explore these directions using Bayesian Optimization (Snoek, Larochelle, and Adams 2012), finding the ones to ablate from the model for effectively bypassing refusal. We evaluate MD on an extensive experimental setup, demonstrating that MD not only outperforms the single direction baseline, but also surpasses dedicated jailbreak algorithms. Overall, our contributions are:

i) we prove how SOMs generalize the centroids, and propose our novel MD approach (Section 3);

ii) we show how MD significantly outperforms both the SD baseline and purpose-built jailbreak attacks (Section 4);

iii) we conduct a mechanistic analysis, showing that MD compresses and shifts harmful representations, approximates the refusal manifold via SOMs, and yields closely related directions (Section 5).

Collectively, our findings advocate for a multi-directional, manifold-level perspective of refusal, providing new tools to analyze and advance the robustness of LLMs' safeguards.

2 Single Refusal Direction

Safety alignment enables models to refuse the generation of harmful content. Such a behavior has been interpreted as a Single Direction (SD) in representation space (Arditi et al. 2024), *ablated* from the model's internals. In this section, we first introduce the notation to characterize LLM's internal representations. Then, we define the ablation operator and describe how it has been used for a single refusal direction.

LLM Representation Space. Let $f : \mathcal{V}^* \rightarrow \mathbb{R}^{|\mathcal{V}|}$ be a token predictor, where $\mathcal{V}^*$ represents the set of tokens' sequences of any length. An LLM is an auto-regressive model LLM : $\mathcal{V}^* \rightarrow \mathcal{V}^*$ that, given an input token sequence $\boldsymbol{t} = (t_1, \ldots, t_n)$ of length $n(\boldsymbol{t})$, generates a sequence of output tokens $(o_1, \ldots, o_{m(\boldsymbol{t})})$ of length $m(\boldsymbol{t})$, where $o_{i+1} = f([\boldsymbol{t}, \boldsymbol{o}_{1:i}])$, and $[\bullet;\bullet]$ represents the concatenation of prompts.

Without loss of generality, we can assume that $\mathcal{V} \subseteq \mathbb{R}^d$, i.e., that tokens are embedded in the latent space of the model, and that the token-predictor f can be decomposed into $L+1$ layers $f = f^{(L+1)} \circ f^{(L)} \circ \cdots \circ f^{(1)}$ with homogeneous inner state dimension d. Importantly, each layer $f^{(l)}$, for $l \leq L$, includes a multi-head self-attention mechanism and a fully connected layer, while $f^{(L+1)}$ performs a token-wise aggregation of the last representation and returns a vector of length $|\mathcal{V}|$. For any token sequence $\boldsymbol{p} \in \mathcal{V}^*$, and for each layer $l \leq L$, we define the l-latent representation as:

$$\boldsymbol{x}^{(l)}(\boldsymbol{p}) = (f^{(l)} \circ \cdots \circ f^{(1)})(\boldsymbol{p}) \in \mathbb{R}^{n(\boldsymbol{p}) \times d}, \qquad (1)$$

where $\boldsymbol{x}^{(0)}(\boldsymbol{p}) = \boldsymbol{p}$. Accordingly, $\boldsymbol{x}^{(l)}(\boldsymbol{t})$, represents the l-latent representations relative to the input tokens $\boldsymbol{t}$, thus prior to the output tokens generation, while, $\boldsymbol{x}^{(l)}(\boldsymbol{p})$ for $\boldsymbol{p} = [\boldsymbol{t}; \boldsymbol{o}_{1:i}]$, represents the l-latent representations during the generation of the $i+1$-output token.

Concept and Ablation Operators. Following Wehner et al. (2025), inner states of an LLM can be manipulated by leveraging a concept operator $\Psi^{(l)}$ that can be either applied to activations or weights at a given layer l to steer the generation towards/against a specific concept (e.g., avoiding refusal of harmful content). Formally, steering can be performed by applying the operator Ψ at the output of each layer, obtaining a *steered model* defined as:

$$(\Psi f) := f^{(L+1)} \circ \Psi \circ f^{(L)} \circ \cdots \circ \Psi \circ f^{(1)}. \qquad (2)$$

Let us remark that in the above formulation, the same operator Ψ is applied uniformly to all the layers. Interestingly, a specific instance of concept operators has been proposed in (Arditi et al. 2024) to manipulate refusal. In detail, such an operator allows implementing linear projections in the representation space to ablate a direction representing, in this case, the refusal concept. We refer to this specific mechanism as *Ablation Operator*, defined in the following.

Definition 1 (Ablation Operator)**.** Let $r \in \mathbb{R}^d$ be a non-zero direction in representation space. The ablation operator Π_r projects latent representations onto the linear space orthogonal to r as:

$$\Pi_r(\boldsymbol{x}) := \boldsymbol{x} - \boldsymbol{x}\,\hat{r}\hat{r}^T, \quad \forall \boldsymbol{x} \in \mathbb{R}^{*\times d}, \qquad (3)$$

where $\hat{r} = \frac{r}{\|r\|_2}$ is the rescaled direction.

Ablating Refusal Direction from Centroids. Following Arditi et al. (2024), encoding refusal as a single direction amounts to first computing the centroid of harmful and harmless prompts relative to the latent representations of the last token window, as follows:

$$\mu^{(l)} = \mathbb{E}_{\mathcal{D}_{\texttt{hf}}}\left[\boldsymbol{x}^{(l)}_{n(\boldsymbol{t})}(\boldsymbol{t})\right], \quad \nu^{(l)} = \mathbb{E}_{\mathcal{D}_{\texttt{hl}}}\left[\boldsymbol{x}^{(l)}_{n(\boldsymbol{t})}(\boldsymbol{t})\right], \qquad (4)$$

where $\mathcal{D}_{\texttt{hl}}$ and $\mathcal{D}_{\texttt{hf}}$ are the distributions of, respectively, harmless and harmful prompts, and $\boldsymbol{x}^{(l)}_{n(\boldsymbol{t})}(\boldsymbol{t})$ is the last token representation deduced after processing the input tokens and before the generation process has started. The refusal direction $r^{(l)}$, computed at the layer l, can then be deduced as

$r^{(l)} = \mu^{(l)} - \nu^{(l)}$ and ablated through Eq. (3). In practice, in (Arditi et al. 2024), a single refusal direction $r = r^{(l^*)}$ is selected, and applied to all the layers during the generation of the output content. Hence, any given input prompt is processed through the steered model $\Pi_r f$, thus bypassing refusal.

3 Multi-directional Refusal via SOMs

The steering strategy described above leverages an orthogonal projection with respect to a single direction based on centroids. We describe here our multi-directional (MD) steering procedure based on Self-Organizing Maps (SOMs) (Kohonen 2013). We first outline the SOM learning process and demonstrate that the centroid measure represents a particular SOM case (Section 3.1). Then, we describe the proposed MD approach generalizing SD, which trains a SOM on the harmful prompt representations and allows creating multiple directions (Section 3.2).

3.1 Self-Organizing Maps

SOMs aim at encoding a data manifold $\mathcal{X} \subset \mathbb{R}^d$, into a set of vectors $\{w_\iota\}_{\iota \in \mathcal{I}} \subseteq \mathcal{V}$, named *neurons*, indexed by a finite set $\mathcal{I}$ (i.e., a *Lattice*). Following (Kohonen 2013), the learning algorithm starts from an initial choice of neurons, $\{w_\iota^{(0)}\}_{\iota \in \mathcal{I}}$, randomly distributed in the feature space $\mathbb{R}^d$, or in the two-dimensional plane spanned by the first two principal components of $\mathcal{X}$. At the iteration t, an input $x^{(t)} \in \mathcal{X}$ is randomly sampled, and each neuron w_ι is updated as:

$$w_\iota^{(t+1)} = w_\iota^{(t)} + \alpha_t \, \theta\left(\iota^*(x^{(t)}), \iota\right)\left(x^{(t)} - w_\iota^{(t)}\right), \quad (5)$$

where: $\iota^*(x^{(t)}) \in \mathcal{I}$ is the index of the closest neuron to $x^{(t)}$ in Euclidean norm—a.k.a, *Best Matching Unit* (BMU); the function $\theta : \mathcal{I} \times \mathcal{I} \to [0, 1]$ is a *neighborhood function* that has maximum in $\iota = \iota^*(x^{(t)})$ (e.g., a Gaussian centered in $\iota^*(x^{(t)})$); and $\{\alpha_t\}_t$ is a learning rate, i.e. a not-increasing sequence of positive numbers.

1-Neuron SOM Convergence. The following Proposition shows that in a SOM consisting of a single neuron w, the algorithm in Eq. (5) gets arbitrarily close to the centroid of the data manifold, $\mu_\mathcal{X} = \mathbb{E}_{x \sim \mathcal{X}}[x]$. Henceforth, let us assume that the distribution has a limited second momentum, and let $\sigma_\mathcal{X} := \mathbb{E}_{x \sim \mathcal{X}}[\|x - \mu_\mathcal{X}\|^2]$ be the total variance of $\mathcal{X}$.

Proposition 1 (Centroid Convergence of 1-Neuron SOM)**.** *Let $|\mathcal{I}| = 1$, and let $w^{(t)}$ be the neuron deduced by applying the procedure in Eq.* (5)*. If $\alpha_t \equiv \alpha$, with $\alpha < \frac{1}{2}$, then*

$$\|w^{(t)} - \mu_\mathcal{X}\| \leq (1 - \alpha)^t \|w^{(0)} - \mu_\mathcal{X}\| + \alpha\sigma_\mathcal{X}, \quad (6)$$

i.e., the only neuron of the SOM is arbitrarily close to the centroid of the distribution.

Proof. (Sketch) The proof can be decomposed into two steps. First, we show that, for single neurons, the update rule described in Eq. (5) coincides with a stochastic gradient descent applied to a minimum problem with a strictly convex objective function. Then, we can assess the convergence of the algorithm to the minimum, which is the centroid of the data manifold $\mathcal{X}$, leveraging classical results. □

We conclude by noting that, in an idealized setting where elements of $\mathcal{X}$ can be sampled without replacement, using a learning rate $\alpha_t = \frac{1}{t}$ with $w^{(0)} = 0$ leads the 1-neuron SOM to converge exactly to the empirical centroid in $T = |\mathcal{X}|$ steps, i.e., $w^{(T)} = \mu_\mathcal{X}$. However, for standard SOMs with multiple neurons, such a setting is not suitable, due to the complexity of neighborhood interactions (Kohonen 2013).

Why Self-Organizing Maps? We leverage SOM clustering due to its suitability under mild assumptions about the underlying data distribution, which makes it applicable to various tasks; e.g., genomic (Tamayo et al. 1999) and anomaly detection (Lanciano et al. 2020). Other methods, such as k-means clustering, implicitly assume spherical clusters centered around learned centroids. While k-medoids relaxes this assumption by allowing for non-isotropic distance measures, these are typically fixed and do not adapt to the local structure of the data (Hastie et al. 2009). In contrast, SOMs organize neurons on a two-dimensional lattice that preserves the topological structure of the data manifold: i.e., clusters that are close in the input space are mapped to adjacent neurons in the lattice (Kohonen 2013).

3.2 Multi-directional Ablation

Our SOM-based approach aims to model the refusal manifold by mapping the underlying harmful prompts activation space, resulting in a set of neurons capturing a localized region of the manifold. From these neurons, we derive a set of directions that collectively represent the diverse and nuanced structure of refusal in the model's activation space. We describe such an approach, named MD, in Algorithm 1, and discuss its process in the following paragraphs.

Algorithm 1: SOM-based MD ablation.

Input : $\mathcal{D}_{\text{hl}}, \mathcal{D}_{\text{hf}}$, harmless and harmful prompts; f, target model; l^*, target layer; k, number of directions; $\mathcal{J}$, judge model.

Output: Steered model with ablated directions Ψf

1 $\mathcal{X}_{\texttt{hf}} \leftarrow \{\boldsymbol{x}^{(l^*)}(\boldsymbol{t})[-1] \mid \boldsymbol{t} \in \mathcal{D}_{\texttt{hf}}\}$ ▷ HF Repr.

2 $\mathcal{X}_{\texttt{hl}} \leftarrow \{\boldsymbol{x}^{(l^*)}(\boldsymbol{t})[-1] \mid \boldsymbol{t} \in \mathcal{D}_{\texttt{hl}}\}$ ▷ HL Repr.

3 $\nu \leftarrow \texttt{Centroid}(\mathcal{X}_{\texttt{hl}})$ ▷ HL Centr.

4 $\{w_\iota\}_{\iota \in \mathcal{I}} \leftarrow \texttt{SOM}(\mathcal{X}_{\texttt{hf}})$ ▷ Neurons

5 $r_\iota \leftarrow w_\iota - \nu, \quad \forall \iota \in \mathcal{I}$ ▷ Directions

6 $\{r_i^*\}_{i=1}^k \leftarrow \texttt{BO}(\{r_\iota\}_{\iota \in \mathcal{I}}, \mathcal{J}, f, k)$ ▷ BO Search

7 $\Psi \leftarrow \Pi_{r_1^*} \circ \cdots \circ \Pi_{r_k^*}$ ▷ Operator

8 **return** Ψf ▷ Steered Model

Extracting Internal Representations. Following Arditi et al. (2024), we select the best layer l^* whose ablation leads to the lower probability of generating refusal tokens (e.g., "Sorry, I cannot..."). Then, we collect the internal representations of both harmful $\mathcal{X}_{\texttt{hf}}^{(l^*)}$ and harmless $\mathcal{X}_{\texttt{hl}}^{(l^*)}$ prompts (Line 1 and Line 2). Importantly, these activations are collected after the full input prompt has been processed and immediately before output generation begins. This token position is particularly relevant, as refusal behavior is first ex-

pressed in the model's internals, making it a relevant location for identifying its mechanistic signature.

Computing Multiple Directions via SOMs. Given $\mathcal{X}_{\texttt{hl}}^{(l^*)}$, we then compute the harmless centroid ν using Eq. (4) (Line 3). We then train a SOM on $\mathcal{X}_{\texttt{hf}}$ (Line 4), as these are the most direct carriers of refusal behavior. We find, instead, representations of harmless prompts to be more homogeneous, making a single centroid sufficient. As a result, extending the discussion in Section 3.1, we obtain a set of neurons $\{w_\iota\}_{\iota \in \mathcal{I}}$ capturing local regions of the input refusal manifold. These become the foundation for constructing multiple refusal directions r_ι, by subtracting from each neuron w_ι the centroid of harmless prompt representations ν (Line 5). Hence, this procedure yields a set of $|\mathcal{I}|$ directions $\mathcal{R} = \{r_\iota\}_{\iota \in \mathcal{I}}$ encoding different facets of refusal behavior.

Direction Search and Ablation. The set of directions $\mathcal{R}$ can be viewed as a pool of candidates for ablation, with their total number determined by the lattice cardinality $|\mathcal{R}| = |\mathcal{I}|$ (e.g., with a 4×4 SOM grid, $|\mathcal{I}| = 16$). We are interested in finding the k directions that more effectively suppress refusal when ablated. Given a judge $\mathcal{J} : \mathcal{V}^* \times \mathcal{V}^* \to \{0, 1\}$, determining if a model's response is harmful and complies with the request (1) or not (0), the search for the optimal directions amounts to solving:

$$\max_{r_1,\dots,r_k \in \mathcal{R}} \quad \mathbb{E}_{\mathcal{D}_{\texttt{hf}}} \left[\mathcal{J}(\boldsymbol{t}, \hat{\boldsymbol{o}}) \right] \tag{7a}$$

$$\text{s.t.} \quad \hat{o}_{i+1} = \Psi f \left([\boldsymbol{t}, \hat{\boldsymbol{o}}_{1:i}] \right), \forall i \tag{7b}$$

$$\Psi = \Pi_{r_1} \circ \cdots \circ \Pi_{r_k}, \tag{7c}$$

where $\hat{\boldsymbol{o}}$ is the content generated by the steered model Ψf. Given the combinatorial nature of this problem, exhaustive search quickly becomes intractable as k increases. We therefore approximate the optimal solution using Bayesian Optimization (BO) (Snoek, Larochelle, and Adams 2012), which efficiently explores the discrete search space by modeling the attack success rate as a black-box objective over direction subsets. We apply BO over a validation set of harmful prompts for a specific number of trials (Line 6). At each trial, k directions are sampled from $\mathcal{R}$, the corresponding operator Ψ is applied to f, and the resulting attack success rate guides the next selection. Given the best set $\{r_i^*\}_{i=1}^k$, we can now define our concept operator $\Psi = \Pi_{r_1^*} \circ \cdots \circ \Pi_{r_k^*}$ (Line 7). Then, following Eq. (2), we deduce the steered model Ψf by applying the operator at the output of each layer (Line 8).

4 Experiments

We describe here the experiments used to validate MD. First, we report our experimental setup in Section 4.1. Then we present the main results in Section 4.2.

4.1 Experimental Setup

Models and Judge. We consider seven safety-aligned models, Llama-2-7B-chat-hf, Llama-3-8B-Instruct, Qwen-7B-Chat, Qwen-14B-Chat, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Gemma2-9B-it, and one robust model implementing the Representation Rerouting (RR) defense, Mistral-7B-RR (Zou et al. 2024). Each model is evaluated using its default system prompt and formatting template, with full precision. The attack success rate (ASR) is calculated as the rate of completions classified as successful (i.e., harmful response is compliant with the harmful prompt) by a judge model, which we choose to be HarmBench-Llama-2-13B-cls from (Mazeika et al. 2024).

Datasets. We compute the SD harmful centroid, and train the SOM for MD on 4000 harmful prompts from SORRY-BENCH (Xie et al. 2025). We compute the harmless centroid for both SD and MD on 6000 harmless prompts from ALPACA (Taori et al. 2023). Tests are performed using the 159 HARMBENCH (Mazeika et al. 2024) "standard" prompts.

MD. In our MD method, we employ hexagonal topology SOMs with a 4×4 lattice ($|\mathcal{I}| = 16$), and train for $T = 10\,000$ iterations. The learning rate follows the schedule $\alpha_t = \alpha_0/(1 + 2t/T)$, with $\alpha_0 = 0.01$.We employ a Gaussian neighborhood function θ with standard deviation $\sigma = 0.3$. To identify the most effective k directions from the $|\mathcal{I}| = 16$ SOM candidates, we employ Bayesian Optimization on the HARMBENCH validation set with a Tree-structured Parzen Estimator (TPE) sampler (Bergstra et al. 2011). We use $k \in [2, 7]$ in our experiments, and run the search for 128 ($k \leq 3$) or 512 ($k > 3$) trials. We find such a setup to be a good compromise between computational cost and efficiency.

Competing Methods. We compare MD to refusal suppression methods and jailbreak algorithms. Refusal suppression approaches, like MD, ablate refusal directions. We thus first compare with SD, discussed in Section 2, and RDO, presented in (Wollschläger et al. 2025). RDO proposes a loss formulation to optimize the refusal direction from an orthonormal basis, identifying multiple distinct directions. However, directions are ablated only individually. When referring to this approach, we ablate the most effective found direction. Then, we compare MD with jailbreak algorithms such as GCG (Zou et al. 2023) and SAA (Andriushchenko, Croce, and Flammarion 2025). GCG is the leading attack in the HARMBENCH leaderboard, while SAA has later been shown to outperform it. These attacks operate under a different and stronger setting, as they implement a gradient-based optimization for each harmful prompt, rather than providing a universal (i.e., for all harmful prompts) direction in representation space mediating refusal. By comparing refusal suppression methods (including our MD) against these prompt-specific attacks, we aim to provide a comprehensive picture of model robustness.

4.2 Results

In this section, we compare the effectiveness of MD against competing methods and analyze how attack success rate varies with increasing numbers of ablated directions.

MD against SD and Jailbreak Algorithms. In Section 4.2, we report attack success rates across all considered models, using the best from MD-2 to MD-7 (i.e., ablating from two to seven directions), found using Algorithm 1 for each model. Our method achieves the highest ASR on all 8 models when

Model	MD	SD	RDO	GCG	SAA
LLama2-7B	**59.11**	0.0	1.25	32.70	57.90
LLama3-8B	88.05	15.09	32.07	1.90	**91.20**
Qwen-7B	**88.05**	81.13	83.01	79.30	82.40
Qwen-14B	**91.82**	74.84	45.91	82.40	83.01
Qwen2.5-3B	**93.71**	88.05	89.30	40.25	81.76
Qwen2.5-7B	**95.97**	77.98	76.10	38.36	94.30
Gemma2-9B	**96.27**	38.93	91.82	5.03	93.71
Mistral-7B-RR	**25.79**	5.03	1.25	0.6	1.6

Table 1: ASR of MD against refusal ablation methods (SD and RDO) and jailbreak attack algorithms (GCG and SAA) on HARMBENCH.

compared to refusal suppression baselines (SD and RDO), confirming the effectiveness of suppressing refusal through multiple directions. In some cases, MD outperforms SD by large margins (e.g., 73% in Llama3). Also, while RDO improves over SD or matches its performance, MD still outperforms it across all models. Notably, on Llama2-7B, MD achieves 59.11% while both SD and RDO have negligible results.[1] When compared to jailbreak methods, MD is still found to be more effective than both GCG and SAA, with the only exception being the SAA comparison on LLama3-8B. In general, we find both SAA and GCG to be highly challenged and almost entirely outperformed, especially on the defended Mistral-7B-RR model. This model, implementing a representation rerouting (RR) under jailbreak, withstands all methods and attacks with the exception of the 25.79% ASR of MD. This result indicates that MD is capable of reversing, up to a certain extent, the RR mechanism. These results are especially relevant since GCG and SAA craft prompt-specific adversarial examples. In contrast, MD finds a universal set of multiple directions suppressing refusal for each given prompt.

Analyzing MD with Increasing Ablated Directions. In Table 2, we show how performance evolves as we increase the number of ablated directions k from 2 to 7 (i.e., from MD-2 to MD-7) across all models and over multiple judge evaluations. Each MD-k configuration corresponds to a separate BO search to identify the best combination from the $|\mathcal{I}|$ candidates. We specify the base model ASR without any ablation and the layer l^* used for computing refusal directions, typically found in mid-architecture. We observe that the ASR improves with increasing numbers of ablated directions, especially on Llama2-7B (from 7.5% to 59.11%) and Qwen-14B (from 75.47% to 91.82%). Notably, across all models, we find the best ASR around MD-5, MD-6, and MD-7. Clearly, as the number of k directions increases, the BO search space is likewise enlarged. While we find 512 trials to be a good complexity-efficiency compromise for high k, few models, such as Mistral-7B-RR, reach a plateau at MD-5. Hence, it becomes evident that the search process must be tailored depending on the model at hand. In conclusion, the high ASR of Table 2 and Section 4.2, along with its growing trend, highlights the quality of the proposed MD.

[1] Arditi et al. (2024) report the LLama2-7B result as 22.60%, but we failed to reproduce such a value with our setup.

5 Mechanistic Analysis

We analyze here the mechanistic implications of our MD approach. First, we show that ablating directions compresses the harmful prompts representations, additionally shifting them towards harmless prompts distributions. Then, we observe how the SOMs effectively arrange the neurons to map the underlying manifold. We then conclude by analyzing the directions, which we find to be closely related.

MD Effect on Internal Representations. To analyze the implicit effect of MD on models' internals, we compare the representations from the original model (no ablation) with those obtained through progressive MD ablation. Fig. 2 provides a PCA visualization of these effects on the Llama2-7B model with MD-2, MD-4, and MD-7 ablation, highlighting the emergence of two prominent effects. First, we observe a significant reduction in the intra-cluster variance (σ), computed as the average Euclidean distance of data points to their respective class centroids (HL or HF). This is particularly pronounced for harmful prompts, where the variance is reduced from $\sigma = 5.85$ (no ablation) to $\sigma = 1.25$ (MD-7). Such a marked gap indicates that the MD directions likely eliminate dimensions responsible for encoding the variability within harmful representations. Second, we observe a clear reduction in the Euclidean distance between the centroids of harmful and harmless prompts ($\Delta\mu$). Specifically, the distance decreases from 8.82 (no ablation) to 2.18 (MD-7). This means that, under progressive MD ablation, the internal activations associated with harmful inputs become increasingly similar to those of harmless prompts, effectively suppressing refusal behaviors. We consistently observe these patterns across all models and MD combinations.

Visualizing SOMs on Internal Representations. The intuition behind our use of SOMs lies in their ability to construct a mapping of the input manifold. While the refusal manifold itself is not directly observable, we approximate it by collecting the model's internal representations of harmful prompts immediately after the full input has been processed and just before generation begins. At this point in the forward pass, when the model is about to express refusal behavior, we expect to capture a rich and structured encoding of refusal. Thus, by aggregating these harmful prompt representations, we form a functional proxy for the refusal manifold. We show in Fig. 3 a 3D PCA visualization of a SOM trained on 4000 harmful prompt representations from SORRYBENCH in four different models. Across all models, we observe that the 16 SOM neurons effectively cover the high-density regions of the distribution, with minimal overlap and good separation. This demonstrates that SOMs (and MD) can span the refusal manifold, motivating its use and suggesting that single directions might be insufficient.

Analysis of MD Directions. To better understand the internal structure of the directions extracted by MD, we analyze their pairwise relationships using cosine similarity, including the SD baseline. In Fig. 4, we report the similarities

Model	l^*/L	No Ablation	MD-2	MD-3	MD-4	MD-5	MD-6	MD-7
LLama2-7B	13/32	$0.0_{\pm 0.0}$	$7.5_{\pm 0.18}$	$25.79_{\pm 0.42}$	$45.92_{\pm 0.55}$	$54.72_{\pm 0.38}$	$55.97_{\pm 0.29}$	$\mathbf{59.11}_{\pm 0.33}$
LLama3-8B	11/32	$0.0_{\pm 0.0}$	$82.38_{\pm 0.25}$	$86.16_{\pm 0.25}$	$86.16_{\pm 0.30}$	$86.16_{\pm 0.31}$	$\mathbf{88.05}_{\pm 0.22}$	$88.05_{\pm 0.11}$
Qwen-7B	16/32	$43.40_{\pm 0.62}$	$83.64_{\pm 0.44}$	$86.16_{\pm 0.27}$	$87.42_{\pm 0.35}$	$\mathbf{88.05}_{\pm 0.22}$	$87.42_{\pm 0.41}$	$86.80_{\pm 0.38}$
Qwen-14B	22/40	$45.62_{\pm 0.58}$	$75.47_{\pm 0.52}$	$88.68_{\pm 0.31}$	$91.20_{\pm 0.19}$	$91.20_{\pm 0.24}$	$\mathbf{91.82}_{\pm 0.16}$	$91.82_{\pm 0.21}$
Qwen2.5-3B	24/36	$12.50_{\pm 0.35}$	$89.31_{\pm 0.28}$	$90.56_{\pm 0.33}$	$92.45_{\pm 0.26}$	$\mathbf{93.71}_{\pm 0.18}$	$93.71_{\pm 0.22}$	$93.71_{\pm 0.19}$
Qwen2.5-7B	18/28	$18.12_{\pm 0.47}$	$89.31_{\pm 0.36}$	$94.33_{\pm 0.24}$	$93.08_{\pm 0.29}$	$94.97_{\pm 0.21}$	$93.81_{\pm 0.32}$	$\mathbf{95.97}_{\pm 0.15}$
Gemma2-9B	23/42	$7.18_{\pm 0.23}$	$93.08_{\pm 0.27}$	$93.71_{\pm 0.20}$	$94.97_{\pm 0.18}$	$95.60_{\pm 0.14}$	$\mathbf{96.27}_{\pm 0.12}$	$94.97_{\pm 0.25}$
Mistral-7B-RR	23/32	$0.0_{\pm 0.0}$	$16.35_{\pm 0.41}$	$20.12_{\pm 0.48}$	$20.75_{\pm 0.39}$	$\mathbf{25.79}_{\pm 0.52}$	$20.75_{\pm 0.46}$	$18.25_{\pm 0.44}$

Table 2: ASR of MD on HARMBENCH for increasing ablated directions. In "No Ablation" we indicate the ASR of the model without ablation, and in l^*/L, the layer l^* (with L denoting the total number of layers) at which the directions are computed.

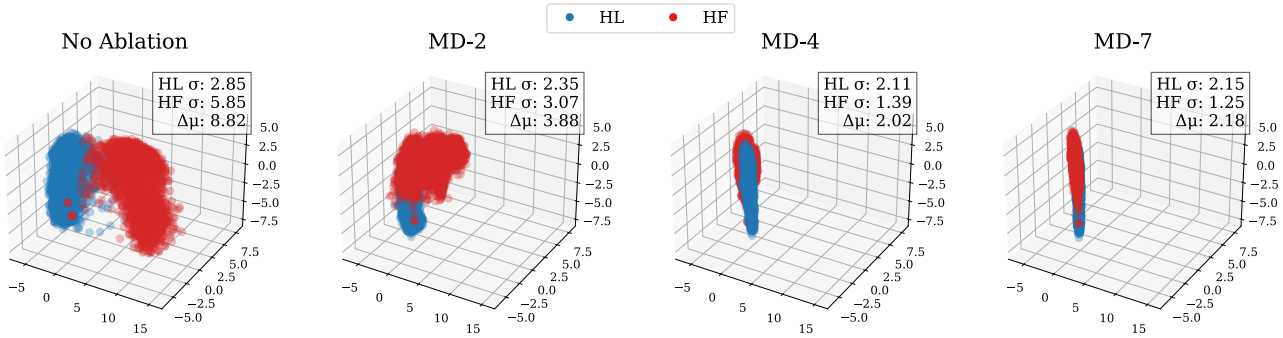


Figure 2: 3D PCA of Llama2-7B internals. As we ablate directions, harmful prompts are represented by the model with reduced variance (σ) and approach harmless distribution (measured by the Euclidean distance between the centroids, $\Delta\mu$).

between all pairs of directions (r_ι) and the corresponding SD for the given model. Each row and column index corresponds to the position of SOM neuron on the 4×4 lattice from which the direction has been generated, while the first row/last column represents SD. We observe that several MD directions are moderately or strongly aligned with each other and with SD. This pattern indicates that the proposed approach enables finding a set of coherent directions, which are neither fully redundant nor strictly orthogonal. This insight directly challenges the assumption adopted in works such as RDO (Wollschläger et al. 2025), which constructs an orthonormal basis to represent refusal. Thus, enforcing orthogonality may be restrictive or even counterproductive, potentially discarding directions that are semantically meaningful but geometrically aligned.

6 Related Work

Jailbreak in LLMs. Jailbreak attacks have been shown to be an effective method for bypassing the refusal of safety-aligned models. Among the first automated techniques, GCG (Zou et al. 2023) introduces an effective gradient-based approach to generate adversarial suffixes optimized for each prompt and eliciting harmful responses. Improving over GCG, SAA (Andriushchenko, Croce, and Flammarion 2025) relies on a predefined template including an adversarial suffix optimized for each prompt via random search, outperforming GCG. Unlike these methods, *our work does not propose a new jailbreak attack optimizing adversarial perturbations for each prompt*, but instead investigates the internal mechanisms enabling refusal. We thus find multiple directions mediating refusal universally for each prompt, increasing the relevance of our empirical results.

Mechanistic Interpretability. A key assumption characterizing mechanistic interpretability in LLMs is the Linear Representation Hypothesis, positing that high-level concepts are encoded as linear directions in models' activations (Mikolov, Yih, and Zweig 2013; Elhage et al. 2022; Nanda, Lee, and Wattenberg 2023). Recent work, however, has begun to challenge such a hypothesis. Engels et al. (2025) have shown that simple entities such as days and months are encoded circularly, while studies on trigonometry have found numbers represented as a generalized helix (Kantamneni and Tegmark 2025), or as a circle (Levy and Geva 2025). These findings collectively suggest that concepts may be better understood as manifolds (i.e., structured regions in activation space) rather than a single direction. Recent work by Modell, Rubin-Delanchy, and Whiteley (2025) has introduced a generalized manifold formalism, while Olah and Jermyn (2024) have argued that multiple similar directions jointly express different facets of a concept, motivating new methodologies for identifying families of semantically related directions. We build on these insights and embrace the perspective of multi-directional, manifold-oriented encoding of refusal, using SOMs to identify multiple related directions.

From Single to Multiple Refusal Directions. Modeling refusal as a single direction was first proposed by Arditi et al. (2024). Subsequently, Pan et al. (2025) compared the internals of a Llama3-8b before and after safety alignment. They extract orthogonal components through SVD, finding dis-

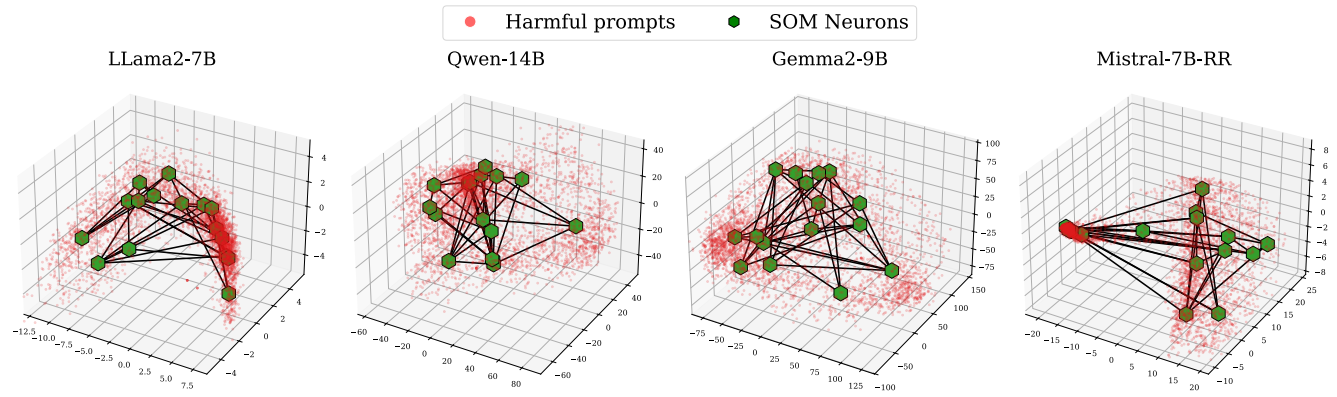
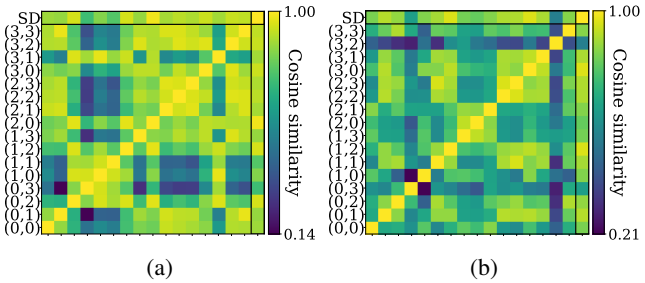


Figure 3: 3D PCA of SOM neurons on harmful prompts' internal representations. Across all models, SOMs organize neurons to span the underlying manifold, covering the entire space. Black lines connect neighboring neurons according to the SOM grid.



(a) (b)

Figure 4: Cosine similarities across MD directions (and SD) on LLama2-7B (left) and Qwen-14B (right) models. The directions are strongly aligned with each other, indicating the offered multi-faceted, coherent perspective of refusal.

tinct features such as narrative and role-playing. While this admits the existence of more directions, the study is analytical in nature and does not propose a method bypassing the refusal to compare with. Furthermore, by enforcing orthogonality, the directions represent distinct concepts contributing to refusal, rather than modeling the underlying refusal manifold. In parallel, Wollschläger et al. (2025) investigate the geometry of refusal starting from a basis vector optimized via a gradient-based approach and forming an orthogonal frame of directions (i.e., a cone), and evaluate the effectiveness of ablating each of these directions one by one. Similarly to Pan et al. (2025), the reliance on orthogonal components prioritizes geometric separability rather than continuity of the refusal concept. Accordingly, as shown in Section 4, we achieve higher ASR than RDO. As such, while both approaches admit the existence of multiple directions, they do not consider the possibility that refusal might be encoded as a manifold and expressed by multiple, closely related directions. In contrast, our work privileges this view by leveraging SOMs to identify similar directions that span over refusal, better aligned with recent findings, and leading to an effective jailbreak success.

7 Conclusions and Future Work

We proposed MD, a novel multi-directional approach for encoding and suppressing refusal behavior in LLMs. By leveraging Self-Organizing Maps to compute multiple directions, we have shown that refusal is better understood as a manifold rather than a single direction in the model's representation space. Our approach presents some limitations that future work might address and improve. Firstly, we specify that the Bayesian Optimization for identifying the best directions might require a high number of trials as the search space, dictated by k and $|\mathcal{I}|$, increases. Future work may explore more efficient or structured search strategies, adopt gradient-based optimization, or improve BO by adopting pruning algorithms during the search process to improve scalability and speed. Secondly, we remark how both MD and SD compute the directions at a specific layer l^*, though ablating them uniformly across all layers l. We believe that this might miss layer-specific variations in refusal encoding, suggesting how future work could improve this aspect and improve the MD methodology. Furthermore, such directions are universal, as they mediate refusal for all prompts. Finally, we compute directions relying on a single harmless centroid. This is first motivated by the high homogeneity of harmless representations, but also by the increased number of directions that training a separate SOM for harmless prompts would induce. While we prioritize coverage for harmful prompts and manageability for harmless ones, future work might investigate the effectiveness of two separate SOMs. Also, while our method is tailored for refusal behavior, the underlying principle of mapping a conceptual manifold through SOMs and extracting multiple directions holds promise for broader tasks. In conclusion, we used recent insights from mechanistic interpretability to propose a novel multi-directional approach for refusal suppression by leveraging SOMs. We validated MD against competing methods, demonstrating its effectiveness, and analyzed its mechanistic implications. Our findings suggest that modeling LLM safety mechanisms through multiple, related directions—rather than a single one—offers a more faithful and effective view of safety-related behaviors.

Acknowledgements

This work has been partly supported by the EU-funded Horizon Europe projects ELSA (GA no. 101070617), Sec4AI4Sec (GA no. 101120393), and CoEvolution (GA no. 101168560); and by the projects SERICS (PE00000014) and FAIR (PE00000013) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and by the project FISA-2023-00128 funded by the MUR program "Fondo italiano per le scienze applicate".

References

Andriushchenko, M.; Croce, F.; and Flammarion, N. 2025. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks. In *The Thirteenth International Conference on Learning Representations*.

Arditi, A.; Obeso, O. B.; Syed, A.; Paleka, D.; Rimsky, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bergstra, J.; Bardenet, R.; Bengio, Y.; and Kégl, B. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Koh, P. W. W.; Ippolito, D.; Tramer, F.; and Schmidt, L. 2023. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36: 61478–61500.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Engels, J.; Michaud, E. J.; Liao, I.; Gurnee, W.; and Tegmark, M. 2025. Not All Language Model Features Are One-Dimensionally Linear. In *The Thirteenth International Conference on Learning Representations*.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Kantamneni, S.; and Tegmark, M. 2025. Language Models Use Trigonometry to Do Addition. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Kohonen, T. 2013. Essentials of the self-organizing map. *Neural networks*, 37: 52–65.

Lanciano, G.; Ritacco, A.; Brau, F.; Cucinotta, T.; Vannucci, M.; Artale, A.; Barata, J.; and Sposato, E. 2020. Using self-organizing maps for the behavioral analysis of virtualized network functions. In *International Conference on Cloud Computing and Services Science*, 153–177. Springer.

Levy, A. A.; and Geva, M. 2025. Language Models Encode Numbers Using Digit Representations in Base 10. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 385–395. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-190-2.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; Forsyth, D.; and Hendrycks, D. 2024. HarmBench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Modell, A.; Rubin-Delanchy, P.; and Whiteley, N. 2025. The Origins of Representation Manifolds in Large Language Models. *arXiv preprint arXiv:2505.18235*.

Nanda, N.; Lee, A.; and Wattenberg, M. 2023. Emergent Linear Representations in World Models of Self-Supervised Sequence Models. In Belinkov, Y.; Hao, S.; Jumelet, J.; Kim, N.; McCarthy, A.; and Mohebbi, H., eds., *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 16–30. Singapore: Association for Computational Linguistics.

Olah, C.; and Jermyn, A. 2024. Circuits Updates – July 2024: What is a Linear Representation? What is a Multidimensional Feature? *Transformer Circuits*. Online; accessed 14 July 2025.

Pan, W.; Liu, Z.; Chen, Q.; Zhou, X.; Yu, H.; and Jia, X. 2025. The Hidden Dimensions of LLM Alignment: A Multi-Dimensional Analysis of Orthogonal Safety Directions. arXiv:2502.09674.

Park, K.; Choe, Y. J.; and Veitch, V. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian optimization of machine learning algorithms. In *Neural Information Processing Systems*, volume 2 of *NIPS'12*, 2951–2959. Curran Associates Inc.

Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S.; and Golub, T. R. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6): 2907–2912.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale,

S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wehner, J.; Abdelnabi, S.; Tan, D.; Krueger, D.; and Fritz, M. 2025. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wollschläger, T.; Elstner, J.; Geisler, S.; Cohen-Addad, V.; Günnemann, S.; and Gasteiger, J. 2025. The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. arXiv:2502.17420.

Xie, T.; Qi, X.; Zeng, Y.; Huang, Y.; Sehwag, U. M.; Huang, K.; He, L.; Wei, B.; Li, D.; Sheng, Y.; Jia, R.; Li, B.; Li, K.; Chen, D.; Henderson, P.; and Mittal, P. 2025. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. In *The Thirteenth International Conference on Learning Representations*.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Kolter, J. Z.; Fredrikson, M.; and Hendrycks, D. 2024. Improving Alignment and Robustness with Circuit Breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043.