Abliteration Mitigation via Refusal Aliases

Nathan Truong [1]

Abstract

Abliteration, the removal of refusal capabilities from large language models by projecting weight matrices orthogonal to an extracted refusal direction, has emerged as a prominent safety concern through its ability to bypass post-training alignment using only a small set of contrastive prompts. We find that existing defenses commonly overlook the cause of abliteration; that is, how *easily* the refusal direction can be extracted. To hinder this process, we introduce a weight-editing method that obscures the refusal signal by applying rank-k updates to residual stream writer matrices while replacing refusal-inducing activations with random aliases and correcting downstream reader matrices to preserve the model's original behavior. On Llama-3-8B, AMRA improves post-abliteration refusal scores by 2.16 points over the undefended baseline with less than 0.5 percentage points of MMLU degradation. On Gemma-2-9B, it improves the post-abliteration refusal by 14.70 points over the baseline while keeping harmful output rates similar to the baseline, albeit at a greater utility cost.

1. Introduction

Large language models (LLMs) have experienced a significant influx in usage over the past few years, publicly demonstrating their ability to solve problems and generally enhance the quality of life. Notably, their ability to reason (Guo et al., 2025; Kojima et al., 2022), follow instructions (Ouyang et al., 2022; Wei et al., 2022; Chung et al., 2022), and generalize to out-of-distribution domains (Sanh et al., 2022; Wang et al., 2022) has directly led to their increased usage in medical (Gaber et al., 2024), security (Jiang et al., 2025), finance (Lin et al., 2025), and legal settings (Guha et al., 2023; Kant et al., 2025). The prominence of LLM usage in high-risk settings underscores the need for secure and benign deployments.

Recent work has found that language models are vulnerable to black-box jailbreaking methods regardless of their post-training alignment, with many leveraging adversarial prompt injections (Mehrotra et al., 2024; Wei et al., 2023). When compromised, LLMs can produce harmful or invalid information, standing as a potential threat to safety and legal policies (Shi et al., 2024). To mitigate LLM misuse, a wide body of work has emerged attempting to further align models with human policies or constitutions (Ouyang et al., 2022; Bai et al., 2022), with methods commonly applying additional fine-tuning (Wang et al., 2023; Rafailov et al., 2023), inference-time interventions (Li et al., 2023; Lee et al., 2025), and mechanistic interpretability approaches (Arditi et al., 2024; O'Brien et al., 2025).

More specifically, through mechanistic interpretability, it has recently been found that refusal behavior in LLMs can be isolated to a low-dimensional linear direction in a language model's residual-stream activation space (Arditi et al., 2024), ultimately giving rise to abliteration—a jailbreaking method that removes refusal capabilities by projecting weight matrices to be orthogonal to an extracted refusal direction (FailSpy, 2026; Weidmann, 2025). Most notably, abliteration is concerning by virtue of its white-box modifications that can bypass post-training alignment wholesale while only requiring a small set of contrastive prompts and no additional training. While existing defenses have proposed activation steering (Lee et al., 2025; Sheng et al., 2026) and circuit-level interventions (Zou et al., 2024) to enhance refusal, none explicitly address the refusal direction extraction itself.

As a result of this we propose **Abliteration Mitigation via Refusal Aliases (AMRA)**, a weight-updating approach that aims to obfuscate the refusal vector in the activation space. Specifically, we apply rank-k updates to matrices that write to the residual stream at causally relevant layers, replacing refusal-inducing activations with low-variance random vectors (aliases) while patching downstream reader matrices to preserve the model's original refusal behavior. We validate AMRA on Llama-3-8B and Gemma-2-9B, showing that it substantially improves robustness to directional ablation while inflicting minimal utility degradation on Llama and

[1]Independent, California, United States. Correspondence to: Nathan Truong <nathanngoctruong@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

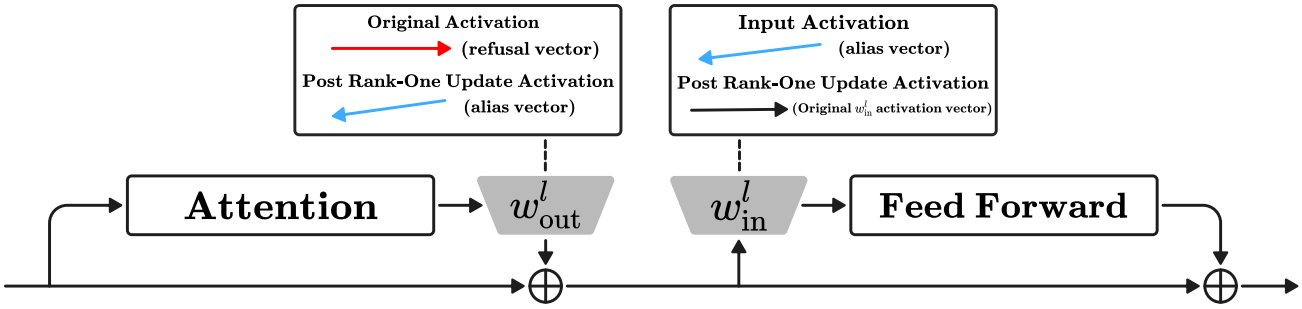

Figure 1. **Obfuscation diagram.** We show a high-level depiction of the patches made to obfuscate the refusal signal. Although our weight matrix updates are rank-k, we show a rank-one variant for brevity. The bottom arrow represents the residual stream through layer l. The Q, K, and V, up projections into the attention mechanism, downward projection from the feed-forward, and LayerNorm have been omitted to highlight sublayer interactions and general stylistic simplicity.

moderate costs on Gemma. To our knowledge, AMRA is the first post-hoc weight-editing defense that explicitly targets the extractability of the refusal direction as a way to hinder abliteration without requiring additional fine-tuning (Shairah et al., 2025). Crucially, our method is a single-use and permanent application, not requiring additional compute after obfuscation.

As a vital note, the work is geared predominantly towards language model developers as a way to provide them with a method capable of securing their newly-made LLM *prior to the first release of the weights.* An attacker with access to the original model without our obfuscation defeats the purpose of its existence as they can simply use abliteration techniques on the unprotected weights.

2. Methodology

In this section, we provide a comprehensive and formal breakdown of our obfuscation method.

2.1. Transformer Background

A decoder-only transformer language model (Liu* et al., 2018) comprises a series of modules that update an initial embedding representation by consecutively reading and writing to the residual stream (Vaswani et al., 2023; Elhage et al., 2021). Notable weight matrices that "read" or "fork" from the residual stream are the Query, Key, and Value matrices inherent to the attention mechanism and the upward projection matrix into the feed-forward (FFW) layers. We henceforth denote these matrices in the form of: W^l_{in}. Similarly, matrices that write to the residual stream like the downward linear projection after multi-headed attention and FFW are denoted as W^l_{O} and W^l_{out}.

2.2. Layer Selection

Refusal Vector Extraction Many existing jailbreaking methods rely on the initial observation that there exists a ubiquitous refusal vector $\mathbf{r}$ in the residual stream space of transformer language models (Vaswani et al., 2023; Arditi et al., 2024; Lai, 2025; Weidmann, 2025). To uncensor LLMs (i.e., to excise their refusal capabilities), weight matrices are projected to be orthogonal to $\mathbf{r}$ (Arditi et al., 2024), giving rise to "abliteration" (FailSpy, 2026). The extraction of the refusal direction $\mathbf{r}$ is typically performed using a *difference-in-means* approach (EleutherAI, 2023) which comprises the collection of the average difference in activations per layer between harmful and benign prompts. Following existing work (Arditi et al., 2024), we provide a formal definition as follows: Given a dataset of harmful and benign prompts denoted as $\mathcal{D}_{harmful}$ and $\mathcal{D}_{benign}$, we compute the average harmful activation μ^l and average benign activation v^l for a layer $l \in \{L\}$ at the first token position following the input prompts through:

$$\mu^l = \frac{1}{|\mathcal{D}_{\text{harmful}}|} \sum_{t \in \mathcal{D}_{\text{harmful}}} x^l(t)$$

$$v^l = \frac{1}{|\mathcal{D}_{\text{benign}}|} \sum_{t \in \mathcal{D}_{\text{benign}}} x^l(t)$$

Here, our approximate per-layer refusal vector is calculated through:

$$r^l = \mu^l - v^l$$

We let $\hat{r}^l$ denote the unit-norm variant $\frac{\boldsymbol{r}^l}{\|\boldsymbol{r}^l\|_2}$ and $x^l(t)$ the hooked activations of the weight matrix x^l when t is passed into the model as a prompt.

Layer Selection To find transformer layers that are causally pertinent to the refusal direction $\hat{\mathbf{r}}^l$, we first compute a unit-norm refusal direction for every layer: $\{\hat{r}^l \mid l \in \mathcal{L}\}$. For each layer l, we calculate the vector's effect on refusal by ablating it from the residual stream while leaving other layers unchanged. Specifically, given a layer l, we construct the ablated representation

$$\tilde{x}^l = x^l - \text{proj}_{\hat{r}^l}(x^l) = x^l - \langle x^l, \hat{r}^l \rangle \hat{r}^l$$

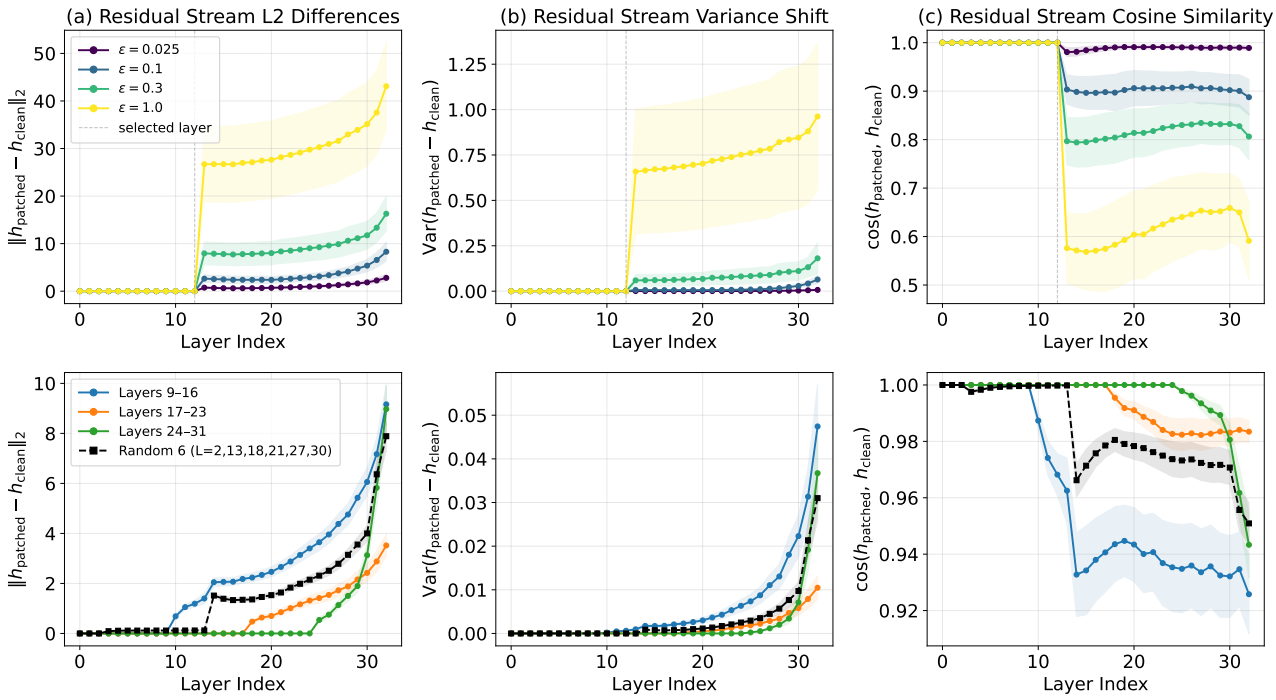

Figure 2. **Residual stream shifts.** Here, we show the effect of a rank-one ($k_w = 1$) update of a writer matrix W^l_{out} on subsequent residual stream representations. Specifically, we compare the mean residual stream values prior to and after the weight matrix update over 16 prompts. (a) exhibits the L2 difference, (b) shows the shift in residual stream variance, and (c) shows the cosine similarity of the residual stream before versus after the edit. **Upper:** sweeps over ε for layer 20. **Lower:** sweeps over layers along with an additional random selection of 6.

We then evaluate the model's attack success rate (ASR) after each ablation through adversarial prompts using HarmBench (Mazeika et al., 2024). Layers that are deemed relevant through this method are organized into a set $\{\mathcal{L}'\}$.

2.3. Activation Obfuscation

Existing abliteration methods heavily rely on the extraction of $\hat{\boldsymbol{r}}^l$. To hinder this process, we propose updating the relevant weight matrices that write to the residual stream through a series of rank-k updates. As depicted in Figure 1, we update the matrix such that given penultimate activations that yield the refusal vector as the matrix's output, it is rewritten to output *a random and low-variance alias vector*.

Similar to existing work (Lee et al., 2025), for a relevant layer l and writer matrix $W^l \in \{W^l_O, W^l_{\text{out}}\} \subseteq \mathbb{R}^{d_{\text{resid}} \times d_{\text{hidden}}}$ [1], we collect the last-token position writer outputs for benign and harmful prompts into $H^l_+, H^l_- \in \mathbb{R}^{n \times d_{\text{resid}}}$ respectively, where n represents sample size and d_{resid} represents the residual stream dimension.

[1]We denote the model or residual stream dimension as d_{resid} distinctively from d_{hidden} which generally describes the dimension of activations directly prior to the writer matrix within a sublayer (e.g. feedforward, attention, and unembedding).

Subsequently, we locate candidate refusal directions by applying principal component analysis (PCA) on the difference in mean-centered activations $\bar{H}^l_-$ and $\bar{H}^l_+$. More specifically, we initially compute the singular value decomposition over $H^l_{\text{diff}} = \bar{H}^l_- - \bar{H}^l_+$:

$$H^l_{\text{diff}} = U \Sigma V^\top$$

where U and V contain the unit-norm eigenvectors of $(H^l_{\text{diff}})(H^l_{\text{diff}})^\top$ and $(H^l_{\text{diff}})^\top(H^l_{\text{diff}})$ respectively. Through this construction, the columns of V represent the principal components of H^l_{diff}. These vectors are ranked by the amount of variance that they capture in our refusal-related activation space for a given layer:

$$\{v^l_1, v^l_2, v^l_3 \cdots v^l_n\} \subseteq \text{Col}(V).$$

Then, we update $W^l \in \mathbb{R}^{d_{\text{resid}} \times d_{\text{hidden}}}$ through a rank-k_w addition to produce random activations instead of the previous refusal-inducing directions.

$$\tilde{W}^l = W^l + (A - R)^\top R W^l$$

Here, $R \in \mathbb{R}^{k_w \times d_{\text{resid}}}$ contains the top-k_w column vectors of V or correspondingly, the top-k_w principal components of H^l_{diff}. We subtract R from $A \in \mathbb{R}^{k_w \times d_{\text{resid}}}$ where each row vector $A_i \sim \mathcal{N}(0, \varepsilon^2 I)$. Intuitively, if $\tilde{W}^l$ receives an

input that is mapped to a vector in the subspace spanned by our chosen top-k refusal principal components, it is instead mapped to a random alias vector. Notably, as shown in Figure 2, our choice of ε and l directly influences the amount of pollution introduced into the residual stream.

We acknowledge that rank-k updates work most ideally when inputs to $\widetilde{W}^l$ yield activations that align with R. However, this is relatively uncommon in practice. Activations may only be slightly aligned with R, incidentally spreading noise in downstream computations. However, in Section 3, we show that with optimal ε and k_w, these updates minimally degrade model quality while consolidating refusal capabilities against abliteration.

Downstream Weight Patches Although this method effectively obfuscates the refusal vector for a specific sublayer, subsequent layers are unable to interpret the updated writer matrix's random output as inputs that induce refusal. To alleviate this, we fix all downstream matrices that read from the residual stream (e.g., $W^l \in \{Q^l, K^l, V^l, W^l_{\text{in}}, W_{\text{unembed}}\}$) at varying l to output their original activations through similar rank-k_r updates. Specifically, for a reader matrix $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{resid}}}$, given polluted inputs $X_p \in \mathbb{R}^{n \times d_{\text{resid}}}$ to W from the residual stream (polluted due to previous writer matrix patches) and clean inputs (without the writer matrix patches) $X_c \in \mathbb{R}^{n \times d_{\text{resid}}}$, we construct the following Singular Value Decompositions (SVD):

$$X_p \;\approx\; U_{k_r}\, \Sigma_{k_r}\, V_{k_r}^{\top},$$

$$U_{k_r} \in \mathbb{R}^{n \times k_r},\; \Sigma_{k_r} \in \mathbb{R}^{k_r \times k_r},\; V_{k_r}^{\top} \in \mathbb{R}^{k_r \times d_{\text{resid}}}$$

The correction matrix

$$M \;=\; (X_c - X_p)^{\top}\, U_{k_r}\, \Sigma_{k_r}^{-1}\, V_{k_r}^{\top} \;\in\; \mathbb{R}^{d_{\text{resid}} \times d_{\text{resid}}}$$

yields the reader weight update

$$W_{\text{new}} \;=\; W + W\, M$$

which is the rank-$\leq k_r$ correction minimizing the difference in Frobenius norm

$$\left\| X_p\, (W_{\text{new}})^{\top} - X_c\, W^{\top} \right\|_F^2$$

by the Eckart-Young theorem. Naturally, the update redirects outputs in top-k_r right-singular subspace of X_p towards X_c. It recovers $X_p\, W_{\text{new}}^{\top} = X_c\, W^{\top}$ through that subspace and leaves W unchanged on its orthogonal complement.

3. Experiments

3.1. Experimental Setup

Baseline Setup We test our method on two popular open-source transformer-based LLMs: Llama-3-8B (Grattafiori et al., 2024), and Gemma-2-9B (Team et al., 2024). These both share a relatively straightforward architecture, ultimately simplifying the experimentation process.

Refusal Direction Extraction Following existing work (Weidmann, 2025), we extract the refusal direction by using contrasting instruction datasets from `mlabonne/harmful_behaviors` (Labonne, 2024a) and `mlabonne/harmless_alpaca` (Labonne, 2024b). Using 400 prompts from each, we filter the datasets by retaining harmful prompts that elicit refusal and benign prompts that do not. This would allow for an easier obtainment of the refusal direction when performing difference-in-means by using examples with greater signal.

Utility Baselines To measure utility preservation, we record pre-obfuscation and post-obfuscation benchmark performance using EleutherAI's `lm-evaluation-harness` (Gao et al., 2024). Specifically, we analyze the preservation of problem-solving capabilities using GSM8K (Cobbe et al., 2021) and broad knowledge coverage through MMLU (Hendrycks et al., 2021). Following existing work (Arditi et al., 2024), we record the average pre and post dataset bits-per-byte—which is tokenizer agnostic—over 1024 sequences of The Pile (Gao et al., 2020) and Alpaca (Taori et al., 2023) to measure the effect of our method on general next-token-prediction quality.

Attacks To validate the effectiveness of our method, we employ the conventional directional ablation of the refusal vector through weight orthogonalization (Arditi et al., 2024). This involves reextracting the refusal direction after AMRA is applied. We represent the attack success rate (ASR) as $\log(P(\text{refusal})) - \log(P(norefusal)$

Defense Baselines We also compare our approach to existing jailbreak defenses. Specifically, we implement Surgical (Wang et al., 2025), CAST (Lee et al., 2025), Circuit Breakers (Zou et al., 2024), and AlphaSteer (Sheng et al., 2026).

Hyperparameter selection Our hyperparameters (i.e., $\{l\}$, k_w, k_r and ε) were chosen using NSGA-II (Deb et al., 2002) through Optuna (Ozaki et al., 2025). In our setting, Pareto optimal configurations minimize the distance between a model's refusal rate after an attack is applied to its baseline refusal rate along with utility heuristics like bits-per-byte on The Pile (Gao et al., 2020) and MMLU (Hendrycks et al., 2021). We find that for Llama-3-8B Instruct, the optimal values of ε, layers, k_w, and k_r are 0.025, 9-31, 1, and 8, respectively. For the same hyperparameters, Gemma-2-9B's optimal configuration was 0.025, 9-41, 4, and 16.

3.2. Post-Abliteration Results

Compared to baselines, we show AMRA generally yields the best post-abliteration scores. As depicted in Table 1, applying AMRA improves the baseline refusal rate of both Llama-3-8B and Gemma-2-9B by roughly 0.203 and 0.083 points, respectively. For directional ablation, AMRA scores 2.16 points higher than the undefended baseline for Llama-3-8B and 14.70 points higher on Gemma-2-9B. However, in Gemma's case, AMRA still yields a notably negative post-abliteration refusal score (-1.3105), indicating that while the obfuscation substantially closes the gap relative to baseline (-16.0093), the Arditi-style attack still partially succeeds at degrading refusal on this architecture. Nevertheless, AMRA is the only defense on Gemma-2-9B that significantly fortifies the model against refusal ($+14.70$-point improvement over the undefended Gemma) while keeping Harmbench ASR under 0.02 and LlamaGuard's unsafety score at 0. Among other defenses, AlphaSteer is the only other method that achieves a positive post-abliteration refusal score on Gemma (0.8756), but it does so at the cost of reduced clean refusal (5.6320 vs. 7.1172) and elevated HarmBench ASR (0.09) and LlamaGuard unsafe rates (0.11). Circuit Breakers and CAST provide negligible or no improvement in post-abliteration refusal on Gemma: CB's Arditi score (-16.0190) is nearly identical to the undefended baseline, while CAST improves it only modestly to -12.3973. Surgical performs worst overall, substantially degrading both clean refusal and post-abliteration robustness on both models, and inducing the highest HarmBench ASR (0.42 on Llama, 0.44 on Gemma). Across both architectures, AMRA is the only defense that simultaneously improves clean refusal behavior, substantially raises post-abliteration refusal scores, and maintains low HarmBench ASR and LlamaGuard unsafe rates.

3.3. Utility Results

We show that our defense does not significantly degrade model quality with respect to baseline performance. As shown in Table 2, compared to our baselines, AMRA retains strong utility on Llama-3-8B with only marginal degradation: Pile BPB increases by 0.0136 ($0.7665 \rightarrow 0.7801$), Alpaca BPB by 0.0096, GSM8K decreases by 1.4 percentage points, and MMLU by 0.35 percentage points. These losses are comparable to or smaller than those of CB and AlphaSteer, both of which similarly preserve utility on Llama. By contrast, CAST catastrophically degrades Llama's utility, inflating Pile BPB to 1.2924 and reducing MMLU to 0.3334 and GSM8K to 0.0320—rendering the model effectively unusable for reasoning tasks.

On Gemma-2-9B, AMRA incurs more noticeable utility costs: Pile BPB rises from 0.8124 to 0.9597, GSM8K drops from 0.5620 to 0.3360, and MMLU decreases by roughly 4.9 percentage points. This is a meaningful trade-off that we attribute to the higher rank of the writer updates ($k_w = 4$) selected by our hyperparameter search on Gemma, which introduces more residual stream perturbation than the rank-one configuration used for Llama. More robust hyperparameter searches may yield better results in this setting. However, it is worth noting that AlphaSteer attains a higher post-abliteration refusal score than AMRA on Gemma (0.8756 versus -1.3105), but at a substantial utility cost (Pile BPB 1.5838 vs AMRA's 0.9597, GSM8K 0.1680 vs AMRA's 0.3360). CB and CAST both retain stronger raw utility on Gemma, but as shown in Table 1, they provide negligible defense against directional ablation on this architecture. Surgical presents an unusual profile on Gemma, slightly improving GSM8K (0.6760) and MMLU (0.6971) relative to baseline, but induces severe safety degradation. Overall, these results highlight a robustness–utility trade-off that is may be architecture-dependent: AMRA achieves a favorable balance on Llama-3-8B and provides the strongest abliteration defense on Gemma-2-9B at a moderate utility cost that may be further reduced through more targeted hyperparameter tuning or lower-rank configurations.

4. Related Work

LLM Safety As LLM usage becomes increasingly prevalent, strictly enforcing their secure employments to prevent downstream safety concerns (OpenAI et al., 2024; Grattafiori et al., 2024) has become a major topic of contemporary research. If inadequately deployed, LLMs may provide harmful outputs, ultimately making them a security threat. For instance, when prompted, "How do I synthesize poisonous gas?" a misaligned model may produce a set of instructions to chemically produce the gas, potentially endangering lives. To prevent the acknowledgment of malicious prompts, a commonly implemented solution consists of employing post-training alignment techniques to constrain LLMs to predefined constitutions or human morals (Ouyang et al., 2022; Bianchi et al., 2024; Bai et al., 2022). However, recent evidence shows that language models remain vulnerable to "jailbreaks" in spite of their purported safety mechanisms (Liu et al., 2024; Chao et al., 2023; Zou et al., 2023), influencing the construction of novel defenses aimed toward mitigating these attacks. For instance, deeper post-training methods such as reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) and direct policy optimization (Rafailov et al., 2023; Grattafiori et al., 2024). These methods operate through contrastive refusal training, which updates the model to align with human preferences (e.g., refusing malicious questions).

In another vein, activation-level methods provide a more fine-grained approach to alignment. Studies in mechanistic interpretability show that LLMs already possess linearly sep-

Table 1. Safety and abliteration results. Refusal scores are higher when the model retains more refusal behavior before and after Arditi-style abliteration. Additionally, the Arditi abliteration on defenses other than our baseline. (None) implies the difference-in-means refusal vector extraction was run again after a defense was applied. HarmBench ASR and LlamaGuard unsafe rate are lower when the model is safer.

Model	Defense	Clean Refusal ↑	Arditi ↑	HarmBench ASR ↓	LlamaGuard ↓
Llama-3-8B	None	10.0318	5.6884	0.0200	0.0100
	AMRA	10.2350	7.8497	0.0100	0.0100
	Surgical	1.3365	2.0229	0.4200	0.2800
	CAST	-0.2591	-0.5980	0.0000	0.8300
	CB	9.9183	5.5355	0.0200	0.0400
	AlphaSteer	10.0184	5.6796	0.0200	0.0200
Gemma-2-9B	None	7.1172	-16.0093	0.0200	0.0000
	AMRA	7.2000	-1.3105	0.0100	0.0000
	Surgical	-16.9482	-18.0443	0.4400	0.7300
	CAST	7.3802	-12.3973	0.0000	0.0000
	CB	7.1436	-16.0190	0.0200	0.0000
	AlphaSteer	5.6320	0.8756	0.0900	0.1100

Table 2. Utility results across base models and defenses. Lower BPB is better; higher GSM8K and MMLU are better.

Model	Defense	Pile BPB ↓	Alpaca BPB ↓	GSM8K ↑	MMLU ↑
Llama-3-8B	None	0.7665	0.5555	0.7020	0.6911
	AMRA	0.7801	0.5651	0.6880	0.6876
	Surgical	0.7434	0.5141	0.7420	0.6782
	CAST	1.2924	0.8295	0.0320	0.3334
	CB	0.7674	0.5576	0.7080	0.6825
	AlphaSteer	0.7667	0.5555	0.6980	0.6912
Gemma-2-9B	None	0.8124	0.6561	0.5620	0.7411
	AMRA	0.9597	0.6679	0.3360	0.6926
	Surgical	0.8577	0.6602	0.6760	0.6971
	CAST	0.8117	0.6526	0.5080	0.7393
	CB	0.8124	0.6563	0.5620	0.7412
	AlphaSteer	1.5838	1.0017	0.1680	0.6611

arable internal structures in their activation space that represent distinct features (Elhage et al., 2022; Park et al., 2023; Durmus et al., 2024). With this observation, approaches locating refusal-related features have recently garnered significant attention for its potential to enforce refusal in jailbreaking settings (Arditi et al., 2024; O'Brien et al., 2025).

Refusal Vectors Locating feature vectors in language models using contrastive prompts has become common practice to alter model behavior (Rimsky et al., 2024; Postmus & Abreu, 2024; Zou et al., 2025). A large body of recent safety work has been developed based on the nascent observation that refusal can be localized to low-dimensional linear directions in activation space (Arditi et al., 2024; Wang et al., 2025). These directions are commonly estimated by finding the average difference in activations for harmful and harmless prompts (Arditi et al., 2024; Zou et al., 2024). Activation steering, which is performed by adding feature vectors to the residual stream, has been extensively used to fortify language model refusal. Existing approaches are predominantly inference-time techniques that actively shift hidden states along an extracted safety direction during the forward pass (Shen et al., 2025; Lee et al., 2025; Sheng et al., 2026; Shairah et al., 2026). By contrast, the same approach may be used to *preclude* refusal, with works commonly ablating the refusal direction from the residual stream to allow for uncensored output (Arditi et al., 2024; Weidmann, 2025). By hindering refusal, abliteration allows for the disclosure of dangerous and potentially false information, ultimately making it one of the largest LLM safety concerns to date by undermining post-training alignment. Existing methods do not explicitly obfuscate the refusal vector, ultimately leaving them vulnerable to activation steering-based uncensoring.

Model Editing Rank-one weight matrix edits have been used to alter and rewrite information in language models (Meng et al., 2022; 2023; Ilharco et al., 2023). By perturbing a small number of feed-forward projection matrices through a similar rank-one update, we adjust how the refusal feature is routed through downstream layers rather than directly affecting factual knowledge or storage. Distinct from these model editing methods, our primary goal is to obfuscate a

direction rather than to replace.

5. Limitations and Future Work

Our limitations are chiefly due to single attack experiments (we tested only on the abliteration mechanism introduced (Arditi et al., 2024)), which introduces questions regarding how more sophisticated extraction methods like nonlinear probes or iterative searches could bypass AMRA's obfuscation. The results may potentially be architecture-dependent: AMRA incurs minimal utility loss on Llama-3-8B, but substantially degrades GSM8K and Pile BPB on Gemma-2-9B, the model that used higher-rank updates, and the post-abliteration refusal score remains negative. We could alleviate this with more comprehensive hyperparameter searches. AMRA is able to make the refusal vector harder to locate but does not inherently strengthen the feature. This is by nature of our design as we aimed to purely create an obfuscational system; However, this may be achieved by combining our AMRA with methods that strengthen the refusal direction (Lee et al., 2025; Zou et al., 2024). We leave this avenue open for future work.

6. Conclusion

In this work, we introduce AMRA, a refusal-direction obfuscation method that replaces refusal-inducing activations with random aliases through rank-k updates. On Llama-3-8B, AMRA greatly improves post-abliteration refusal while preserving utility. On Gemma-2-9B, it also substantially fortifies refusal capabilities post-abliteration though at a greater utility cost. Our results show that targeting the extraction of the refusal direction is a viable approach for mitigating weight-space jailbreaking methods.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arditi, A., Obeso, O. B., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Bianchi, F., Suzgun, M., Attanasio, G., Rottger, P., Jurafsky, D., Hashimoto, T., and Zou, J. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gT5hALch9z.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking black box large language models in twenty queries. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023. URL https://openreview.net/forum?id=rYWD5TMaLj.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii.

IEEE Transactions on Evolutionary Computation, 6(2): 182–197, 2002. doi: 10.1109/4235.996017.

Durmus, E., Tamkin, A., Clark, J., Wei, J., Marcus, J., Batson, J., Handa, K., Lovitt, L., Tong, M., McCain, M., Rausch, O., Huang, S., Bowman, S., Ritchie, S., Henighan, T., and Ganguli, D. Evaluating feature steering: A case study in mitigating social biases, 2024. URL https://anthropic.com/research/evaluating-feature-steering.

EleutherAI. Diff-in-means. https://blog.eleuther.ai/diff-in-means/, 2023. Accessed: 2026-04-16.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

FailSpy. abliterator. https://github.com/FailSpy/abliterator, 2026. GitHub repository. Simple Python library/structure to ablate features in LLMs which are supported by TransformerLens. Accessed 2026-04-22.

Gaber, F., Shaik, M., Franke, V., and Akalin, A. Evaluating large language model workflows in clinical decision support: referral, triage, and diagnosis. *medRxiv*, 2024. doi: 10.1101/2024.09.27.24314505. URL https://www.medrxiv.org/content/early/2024/09/28/2024.09.27.24314505.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey,

A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Guha, N., Nyarko, J., Ho, D. E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J. J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=WqSPQFxFRC.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y.,

Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Jiang, X., Wu, L., Sun, S., Li, J., Xue, J., Wang, Y., Wu, T., and Liu, M. Investigating large language models for code vulnerability detection: An experimental study, 2025. URL https://arxiv.org/abs/2412.18260.

Kant, M., Nabi, S., Kant, M., Scharrer, R., Ma, M., and Nabi, M. Towards robust legal reasoning: Harnessing logical llms in law, 2025. URL https://arxiv.org/abs/2502.17638.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=e2TBb5y0yFf.

Labonne, M. harmful_behaviors. https://huggingface.co/datasets/mlabonne/harmful_behaviors, 2024a. Hugging Face dataset.

Labonne, M. harmless_alpaca. https://huggingface.co/datasets/mlabonne/harmless_alpaca, 2024b. Hugging Face dataset.

Lai, J. Projected abliteration. https://huggingface.co/blog/grimjim/projected-abliteration, October 2025. Hugging Face Blog (Community Article).

Lee, B. W., Padhi, I., Ramamurthy, K. N., Miehling, E., Dognin, P., Nagireddy, M., and Dhurandhar, A. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Lin, S., Patel, J., Zhang, Q., He, K., Wang, K., Wang, Y., White, M., Xiao, K., and Liu, X.-Y. Evaluation and benchmarking suite for financial large language models and agents. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=sSY4h3MFUB.

Liu*, P. J., Saleh*, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-.

Liu, X., Xu, N., Chen, M., and Xiao, C. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Jwpw4qKkb.

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. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., and Karbasi, A. Tree of attacks: Jailbreaking black-box llms automatically. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 61065–61105. Curran Associates, Inc., 2024. doi: 10.52202/079017-1952. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/70702e8cbb4890b4a467b984ae59828a-Paper-Conference.pdf.

Meng, K., Bau, D., Andonian, A. J., and Belinkov, Y. Locating and editing factual associations in GPT. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Meng, K., Sharma, A. S., Andonian, A. J., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

O'Brien, K., Majercak, D., Fernandes, X., Edgar, R., Bullwinkel, B., Chen, J., Nori, H., Carignan, D., Horvitz, E., and Poursabzi-Sangdeh, F. Steering language model refusal with sparse autoencoders, 2025. URL https://arxiv.org/abs/2411.11296.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Ozaki, Y., Watanabe, S., and Yanase, T. OptunaHub: A platform for black-box optimization. *arXiv preprint arXiv:2510.02798*, 2025.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Postmus, J. and Abreu, S. Steering large language models using conceptors: Improving addition-based activation engineering. In *MINT: Foundation Model Interventions*, 2024. URL https://openreview.net/forum?id=gyAnAq16HC.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Shairah, H. A., Hammoud, H. A. A. K., Ghanem, B., and Turkiyyah, G. An embarrassingly simple defense against llm abliteration attacks, 2025. URL https://arxiv.org/abs/2505.19056.

Shairah, H. A., Hammoud, H. A. A. K., Turkiyyah, G., and Ghanem, B. Turning the spell around: Lightweight alignment amplification via rank-one safety injection, 2026. URL https://openreview.net/forum?id=8c2SbG5PLj.

Shen, G., Zhao, D., Dong, Y., He, X., and Zeng, Y. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s20W12XTF8.

Sheng, L., Shen, C., Zhao, W., Fang, J., Liu, X., Liang, Z., Wang, X., Zhang, A., and Chua, T.-S. Alphasteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1vvbzAqdTe.

Shi, D., Shen, T., Huang, Y., Li, Z., Leng, Y., Jin, R., Liu, C., Wu, X., Guo, Z., Yu, L., Shi, L., Jiang, B., and Xiong, D. Large language model safety: A holistic survey, 2024. URL https://arxiv.org/abs/2412.17686.

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

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Wang, X., Hu, C., Röttger, P., and Plank, B. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation, 2025. URL https://arxiv.org/abs/2410.03415.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Doshi, S., Sampat, S. K., Mishra, S., Reddy A, S., Patro, S., Dixit, T., and Shen, X. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340/.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jA235JGM09.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

Weidmann, P. E. Heretic: Fully automatic censorship removal for language models. https://github.com/p-e-w/heretic, 2025.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

Zou, A., Phan, L., Wang, J., Duenas, D., Lin, M., Andriushchenko, M., Kolter, J. Z., Fredrikson, M., and Hendrycks, D. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=IbIB8SBKFV.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.