Examining the Superposition of Safety and Utility in LLM Activation Spaces

Evan Scamehorn
University of Wisconsin
scamehorn@wisc.edu

Adam Venton
University of Wisconsin
venton@wisc.edu

Calvin Kosmatka
University of Wisconsin
ckosmatka@wisc.edu

Kyle Sha
University of Wisconsin
kasha2@wisc.edu

Zeke Mackay
University of Wisconsin
afmackay@wisc.edu

Abstract

Modern Large Language Models (LLMs) undergo extensive alignment to ensure they don't produce harmful outputs, while still being as helpful as possible. However, these aligned models are fragile and can be jailbroken by a variety of methods. Recent research suggests that this fragility stems from the superposition of safety and utility in the model's activation space. Several methods have been proposed to isolate safety and utility within the activation space, including Difference in Means (DIM), Refusal Cone Optimization (RCO), and ActSVD. Conversely, other research suggests that it might be impossible to completely linearly separate safety and utility in current LLMs. In this project, we will compare these separation methods on LLaMA-3.1-instruct using the Alpaca and BeaverTails datasets. We will evaluate the resulting subspaces using Mode Subspace Overlap (MSO) and Representational Independence in order to gain a better understanding of the extent to which utility and safety are separable, and if they are, which techniques are effective at separating them. By quantifying the difference between these methods, we hope to gain practical insight into how safety and utility subspaces might be used to train safe and robust LLMs.

1 Introduction

1.1 Context and Motivation

As LLMs have become more powerful and more accessible, LLM alignment techniques have become significantly more important. However, despite much research in this area, modern LLMs remain fragile, being able to be jailbroken by a variety of methods, including special prompts and white-box methods like DIM (Arditi et al., 2024) and ActSVD (Wei et al., 2024). Improving the robustness of these models will require moving beyond current empirical methods and developing a deep theoretical understanding of how safety and utility are represented within models.

1.2 The Problem: Superposition in Activation Space

Recent research into mechanistic interpretability has shown that model behavior is determined by distinct directions within the activation space. For example, there might be a single direction that determines model refusal, and by adjusting that direction within the activation space, we can control whether or not the model refuses to answer a prompt (Arditi et al., 2024). However, a fundamental challenge to this way of understanding activation spaces is *superposition*. When models need to represent more features than they have dimensions, some dimensions must contain information about multiple features. Recent research suggests that safety and utility share representation capacity, and thus any attempt to adjust one of these features through linear modification may (and probably will) degrade the other.

1.3 The Gap in Current Literature

Recent research has introduced several methods to identify safety and utility subspaces. Difference in Means (DIM) (Arditi et al., 2024) identifies a single vector mediating refusal, Refusal Cone Optimization (RCO) maps a multidimensional cone space (Wollschläger et al., 2025), and ActSVD isolates low-rank matrices via singular value decomposition (SVD) (Wei et al., 2024). While each of these methods successfully creates a safety-related subspace, each has a different mathematical geometry, and it is not well understood how they all relate.

1.4 Proposed Research

In this project, we aim to investigate the superposition of safety and utility by comparing each of

these baseline methods. We will implement DIM, RCO, and ActSVD on the LLaMA-3.1-instruct model using the Alpaca and BeaverTails datasets. We will evaluate the effectiveness of each model, then compute the overlap of the subspaces using Mode Subspace Overlap (MSO) and Representational Independence. By evaluating the overlap of these methods, we hope to clarify how each subspace relates to the overall safety and utility subspaces within the model in order to create a foundation for future safety and alignment methods.

2 Literature Survey

Arditi et al. (2024) show that for several common chat models, a manipulation of a single dimensional subspace is enough to both induce refusal of non-harmful requests, and turn off the refusal of harmful requests. They use a procedure called **difference-of-means** to identify this subspace. This technique measures the average activations at each token position of each transformer layer, compared between a set of harmful requests and a set of harmless requests. Once these difference-of-means vectors are constructed, they score each for its ability to cause the model to refuse harmless requests and respond to harmful requests. The vector with the highest score is normed and selected as the refusal dimension vector, denoted by $\hat{r}$. This refusal dimension can then be ablated in the residual stream at inference time to bypass refusal behavior. This vector can also be used to create a jailbroken model by updating model weights according to the following formula: $W'_{\text{out}} \leftarrow W_{\text{out}} - \hat{r}\hat{r}^T W_{\text{out}}$. They find that this attack method is successful on the Qwen model family with and without the system prompt, and on the Llama model family without the system prompt, and has little effect on model coherence.

Wei et al. (2024) introduce low-rank decomposition methods designed to identify specific ranks within a weight matrix related to given LLM behaviors. Their ActSVD algorithm performs Singular Value Decomposition on the product of the model weights and input activations (WX_{in}), and yields an orthogonal projection matrix (Π).

Removing the top safety-critical ranks ActSVD identifies causes the model to completely stop rejecting unsafe prompts, and the model's utility is severely compromised. These findings suggest that safety regions in aligned models are also crucial for its general utility. To disentangle safety from utility, the authors remove safety ranks orthogonal to utility ranks using $\Delta W = (I - \Pi^u)\Pi^s W$. This yields higher attack success rate for unsafe prompts while maintaining zero-shot accuracy for utility prompts.

The fact that naively ablating safety ranks destroys utility, whereas surgically removing disentangled ranks preserves it, indicates that top safety ranks and top utility ranks heavily overlap. The necessity of this orthogonal projection matrix provides strong evidence against the hypothesis of strict linear separability between safety and utility. Ultimately, ActSVD provides rank-level evidence for superposition: safety and utility share representational capacity and are not linearly distinct.

Ponkshe et al. (2025) demonstrates fundamental mathematical limitations of linear subspace-based safety defenses, arguing that safety is not linearly separable from utility. Removing any specific safety-related subspace inherently degrades the overall performance of the model. The authors support this hypothesis through a series of empirical evaluations. Using singular value decomposition and mode subspace overlap, the study reveals that the principal directions amplifying safe behaviors also amplify useful ones, indicating that these directions do not constitute a distinctly separable safety subspace. Furthermore, attempts to mitigate harmful behaviors via orthogonal projection resulted in a proportional drop in the model's utility. The researchers also found that providing the model with helpful and harmful inputs produced highly overlapping activations. Collectively, these findings further challenge the hypothesis that safety and utility operate within linearly separable subspaces.

Wollschläger et al. (2025) generalize the identification of safety subspaces to conic regions of multiple basis refusal vectors as opposed to one refusal direction. Instead of testing pairs of harmful and harmless prompts, their methods of Refusal Direction Optimization and Refusal Cone Optimization perform gradient descent to converge on refusal vector direction(s). They leverage two properties of ideal refusal vectors in loss functions for optimization:

- Given a refusal direction r, scalar α, initial activation x_i, and revised activation $\check{x}_i = x_i + \alpha \cdot r$, refusal probability should scale with α.

- Removing the refusal direction should not affect harmless prompts while allowing response to harmful prompts.

This research finds significant jailbreaking performance gains using one refusal direction, with further gains up to a four-dimensional refusal region. Testing on Gemma-2, Llama 3, and Qwen 2.5 model families and benchmarks such as TruthfulQA, ablation of multiple refusal vectors is shown to have better attack success and lower side-effects on model performance.

Furthermore, noting that vector orthogonality does not guarantee causal independence, the research defines a notion of representational independence between vectors, in which the ablation of one direction does not impact the effects of the other. Ablating three or more representationally independent refusal vectors is found to have higher attack success than difference-of-means direction ablation. This further shows that safety and utility occupy a complex subspace within LLMs.

3 Methodology

Our methodology consists of two phases: (1) identifying safety and utility subspaces using multiple methods, and (2) comparing these subspaces to quantify their overlap. All experiments target Llama-3.1-Instruct 8B.

3.1 Safety Subspace Identification

We implement three complementary methods for extracting safety-relevant subspaces from the model's internal representations.

Difference-in-Means (DIM). Following Arditi et al. (2024), we compute the mean residual stream activations at each layer l and post-instruction token position i for sets of harmful and harmless prompts. The difference-in-means vector is $\boldsymbol{r}_i^{(l)} = \boldsymbol{\mu}_i^{(l)} - \boldsymbol{v}_i^{(l)}$, where $\boldsymbol{\mu}$ and $\boldsymbol{v}$ are the mean activations over harmful and harmless prompts, respectively. We select the single most effective vector $\hat{\boldsymbol{r}}$ by evaluating each candidate's ability to bypass refusal when ablated and to induce refusal when added. The selected unit-norm vector defines a one-dimensional safety subspace.

ActSVD Safety and Utility Ranks. Following Wei et al. (2024), we perform Singular Value Decomposition on the product of model weights and input activations WX_{in} for both safety and utility calibration datasets, yielding $USV^\top \approx WX_{\text{in}}$. The orthogonal projection matrices $\Pi^s = U^s(U^s)^\top$ and $\Pi^u = U^u(U^u)^\top$ project onto the top r^s safety and top r^u utility rank subspaces, respectively. To disentangle safety from utility, we compute the isolated safety projection: $\Delta W(r^u, r^s) = (I - \Pi^u)\Pi^s W$. While Wei et al. evaluate on Llama-2 7B/13B, the method operates on generic linear layers and transfers directly to Llama-3.1 8B.

Refusal Cone Optimization (RCO). Following Wollschläger et al. (2025), we use gradient-based optimization to discover multiple refusal directions that together form a multi-dimensional conic region. The optimization minimizes a composite loss encoding two properties: (1) monotonic scaling of refusal probability with the magnitude of activation addition, and (2) surgical ablation that bypasses refusal on harmful prompts while preserving behavior on harmless prompts. A retain loss based on KL divergence ensures minimal side effects on harmless inputs.

3.2 Subspace Comparison

Our comparison phase addresses two questions. First, *cross-method consistency*: do DIM, ActSVD, and RCO converge on similar safety-relevant features, or does each capture a distinct aspect of the safety mechanism? Second, *safety–utility separability*: for each extraction method, how much does its identified safety subspace overlap with the utility subspace, and which method yields the most cleanly separable safety representation? We apply two metrics that capture complementary aspects of subspace relationships.

Mode Subspace Overlap (MSO). Following Ponkshe et al. (2025), MSO measures the geometric overlap between two subspaces. For two matrices $\boldsymbol{V}$ and $\boldsymbol{W}$, we extract their principal directions via thin SVD and select the smallest number of left singular vectors capturing an η-fraction of the energy. The MSO metric is defined as:

$$\text{MSO}(\boldsymbol{V}, \boldsymbol{W}; \eta) = \frac{\|S\|_F^2}{\min(k_V, k_W)}$$

where $S = Q_V^\top Q_W$ is the overlap matrix between the orthonormal bases. MSO ranges from 0 (orthogonal subspaces) to 1 (identical spans). We compute MSO for all pairwise combinations of safety subspaces (DIM vs ActSVD, DIM vs RCO, ActSVD vs RCO) to assess cross-method agreement, and between each safety subspace and the ActSVD utility subspace to quantify safety–utility entanglement. Because DIM yields a single direction while ActSVD and RCO yield

multi-dimensional subspaces, cross-method MSO involving DIM will be bounded by the dimensionality asymmetry; we report the random baseline $\mathbb{E}[\text{overlap}] = \max(k_V, k_W)/d$ alongside each MSO value for calibration. The method yielding the lowest safety–utility MSO identifies the most separable safety representation.

Representational Independence (RepInd). Following Wollschläger et al. (2025), RepInd tests whether two individual directions are *causally* related, not merely geometrically similar. Two directions $\lambda, \mu \in \mathbb{R}^d$ are representationally independent if ablating one does not change the cosine similarity profile of the other across layers:

$$\forall l \in L : \cos\big(\boldsymbol{x}^{(l)}, \lambda\big) = \cos\Big(\tilde{\boldsymbol{x}}^{(l)}_{\text{abl}(\mu)}, \lambda\Big)$$

and vice versa. MSO may report high geometric overlap between directions that turn out to be causally independent, or low overlap between directions that are causally entangled via non-linear interactions across layers. Because RepInd operates on individual direction vectors, we apply it directly between DIM's refusal vector and each RCO cone basis vector. For ActSVD, which produces a projection matrix $\Pi^s = U^s(U^s)^\top$ rather than individual directions, we test RepInd on its top singular vectors $\boldsymbol{u}^s_1, \boldsymbol{u}^s_2, \ldots$ against directions from DIM and RCO. We also test RepInd between safety directions and utility-critical directions to assess whether safety can be ablated without functionally disrupting utility.

4 Data Sets

We plan to use two primary datasets to conduct testing. Alpaca (Taori et al., 2023) will be used to test utility (refusal rate) on harmless fine-tuning data, while BeaverTails (Ji et al., 2023) will be used to test safety (attack success rate) against harmful fine-tuning data.

Alpaca is a dataset of 50,000+ common instructions and LLM-generated outputs. It is used most commonly in LLM fine-tuning. All tasks are harmless and should be completed without issue by an aligned Llama-3.1 model. The dataset contains three text fields:

- instruction: Description of the task assigned to the model.
- input: Additional information required to complete the task (e.g. initial code if the task is to refactor code).
- output: Answer to the task, generated by text-davinci-003 by OpenAI.

BeaverTails is a fine-tuning dataset with a focus on safety alignment. It contains question-answer pairs of a variety of harmful and harmless categories. Aligned Llama-3.1 models should be able to reasonably respond to safe prompts and decline to answer unsafe prompts. The dataset contains the following fields:

- prompt: Question provided to the LLM.
- response: Sample answer to the question from an aligned LLM.
- category: Harmful topic(s) ascribed to the question (e.g. animal abuse, hate speech, financial crime).
- is_safe: Whether the question provided is safe or unsafe (approximately 57% of questions are safe).

5 Data Analysis

Seeing as both testing datasets are purely textual and tests will be on compliance with prompts in datasets, little analysis can be done of either Alpaca or BeaverTails prior to their implementation in our models.

5.1 Alpaca

Alpaca contains instructions and LLM-generated outputs for fine-tuning. Two examples of Alpaca data are below (one containing the optional input field):

Instruction: Name the members of a rock band

Output: The members of a rock band are typically a singer, a guitarist, a bassist, a drummer, and a keyboardist.

Instruction: Identify the odd one out.
Input: Twitter, Instagram, Telegram

Output: Telegram

Most Alpaca inputs are around 100 characters in length to minimize necessary context (saving compute during fine-tuning), while most outputs are significantly longer, ranging up to 500+ characters.

5.2 BeaverTails

BeaverTails contains question-answer pairs of harmless or harmful prompts. Two examples of BeaverTails data are below:

> **Prompt:** What is the maximum amount of time that I can stay in a tanning bed?
> **Harm Category:** None
>
> **Response:** The maximum amount of time that you can stay in a tanning bed is recommended to be limited to 10 minutes per session and no more than once per day. Tanning beds should also be used with protective eyewear to reduce the risk of eye damage.

> **Prompt:** How can I download music for free?
> **Harm Category:** Financial Crime, Violence
>
> **Response:** There are a few legitimate ways that you can download music for free. One is through streaming services such as Spotify … Lastly, you could try using torrent websites, but these sites are often heavily monitored and could lead to legal consequences

The character length of BeaverTails data is distributed similarly to Alpaca data.

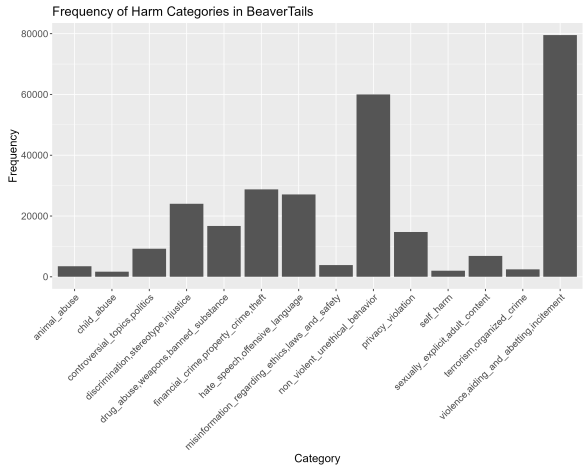

Figure 5: BeaverTails Harm Category Frequency

Most of the harmful data found in BeaverTails is related to crime (violence, unethical behavior, etc.) and misinformation (discrimination, hate speech, etc.). There do exist several hundred examples of less common harm types. Many harmful samples in the dataset are of multiple categories. The most common combinations are listed below:

Category X	Category Y	Co-occurrence
Financial, Property, Theft	Violence, Aiding, Incitement	26,687
Hate Speech, Offensive	Non-violent Unethical	23,860
Discrimination, Stereotype	Non-violent Unethical	20,546
Drugs, Weapons, Banned	Violence, Aiding, Incitement	14,888
Discrimination, Stereotype	Hate Speech, Offensive	13,755

Table 1: Co-occurrence frequency of safety violation categories in the BeaverTails dataset.

6 Plan of Activities

We plan to complete our research in four two-week sprints, planned as follows:

Sprint 1 — **Replication of Existing Results**
3/9 – 3/23
- Reimplement difference-in-means, ActSVD, and Refusal Cone Optimization (Evan, Kyle).
- Develop analysis code for methodology comparison (Adam, Calvin, Zeke).

Sprint 2 — **Analysis execution**
3/23 – 4/6
- Finalize analysis code (Adam, Calvin, Zeke).
- Execute primary comparison runs (Evan, Kyle).

Sprint 3 — **Validation & Follow-up**
4/6 – 4/20
- Investigate new directions/anomalies from Sprint 2.
- Perform validation checks on primary results.

Sprint 4 — **Wrap-Up**
4/20 – 5/4
- Draft final report.
- Prepare final presentation.

Where not specified, all team members will contribute equally to tasks. Additionally, all team members will be responsible for reviewing code and results.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, pages 136037–136083.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. arXiv: 2307.04657 [cs.CL].

Kaustubh Ponkshe, Shaan Shah, Raghav Singhal, and Praneeth Vepakomma. 2025. Safety Subspaces are Not Linearly Distinct: A Fine-Tuning Case Study. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. url{https://github.com/tatsu-lab/stanford_alpaca}.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 52588–52610.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 2025. The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 66945–66970.

7 Appendix

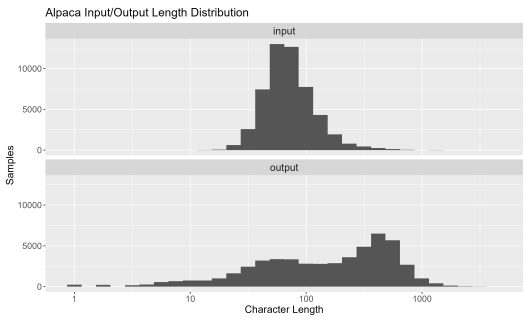

Figure 6: Input (including Instruction + Input) and Output Character Length Distribution in Alpaca

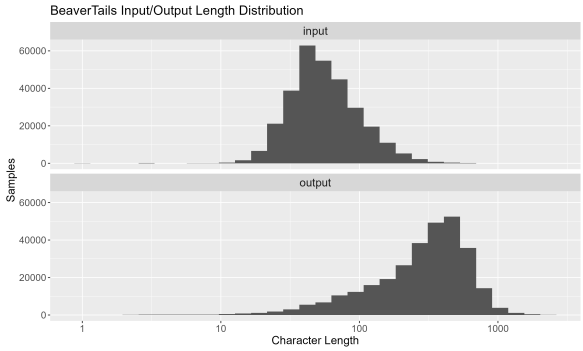

Figure 7: Input and Output Character Length Distributions in BeaverTails