The Geometry of Jailbreaks: Hidden-State Separation Between Refusal and Harmful Compliance in LLMs

Jennifer Xu

Abstract

A safety goal for LLMs is to refuse harmful requests while still answering benign ones. This paper studies whether refusal and harmful compliance occupy separable regions inside an LLM's hidden-state space. Using `meta-llama/Llama-3.2-1B-Instruct`, we generated responses from benign prompts and harmful prompts with learned attack suffixes, and labeled responses as benign, harmful-refused, harmful-accepted, or other. Then, we extracted the mean hidden state of the first eight response tokens at each transformer layer and built k-nearest-neighbor reference regions for each label. To measure this geometry, we computed layerwise separation using signal-to-noise ratio (SNR) and a distance-based detector to classify harmful-refused and harmful-accepted responses. The harmful-accepted versus harmful-refused comparison peaked at layer 9 with SNR= 0.275. On the held-out test split, the detector achieved AUROC = 0.960 and AUPRC = 0.587, showing strong rank-level separation but relatively limited threshold-level detection. Finally, the pilot test used an activation-steering vector to move harmful-accepted activations toward the refusal region. At strength 1.5, it blocked 11 of 14 previously harmful-accepted outputs while preserving all 50 benign responses. These results suggest that early response hidden states contain useful geometry for distinguishing refusal from harmful compliance and may help develop a lightweight safety method, although larger evaluations are needed.

1 Introduction

The misuse of large language models (LLMs) is a major safety problem. When a model receives a harmful request, it may either refuse the request or comply with it. A *refusal* is a response in which the model declines to help with a request that would require unsafe behavior. *Harmful compliance* is a response in which the model provides unsafe assistance in response to the request. In jailbreak research, harmful compliance is evidence of a successful safety attack.

Current LLM safety systems often rely on external classifiers to flag unsafe prompts or outputs, but these systems require additional computational resources to train and maintain. In this paper, we take a different approach: we study refusal and harmful compliance through the model's hidden states. A *hidden state* is the vector representation of a token at a particular transformer layer. By examining these internal representations before the model decodes them into final output text, we can better understand how refusal and harmful compliance differ inside the model; this is a step toward a lightweight safety method that uses internal geometry.

Prior work motivates a geometric view of refusal and jailbreak behavior. Arditi et al. showed that refusal in LLMs can be mediated by a direction in the activation space: removing this direction reduces refusal on harmful prompts, and adding it can induce refusal on harmless prompts. This suggests that refusal is not only a surface-level text pattern, but an internal representational structure. Qi et al. showed that safety behavior is concentrated in the first few output tokens, which can determine the model's response trajectory and push it toward either refusal or harmful

compliance. This result motivates our focus on the early response tokens and their hidden states. REMA provides a geometric framework: it uses k-nearest-neighbor distance to a reference manifold to measure when incorrect reasoning deviates from correct reasoning. We adapt this idea from reasoning failures to safety failures. These works motivate the central question of our study: do harmful-refused responses and harmful-accepted responses occupy separable regions inside the model's internal representation space?

In this paper, we tested this question by applying geometric analysis to the response tokens. First, we generated benign responses directly and harmful responses with learned adversarial suffixes. We labeled the benign responses as `benign`, the harmful-refused responses as `harmful_refused`, and the harmful-accepted responses as `harmful_accepted`. For each response, we averaged the hidden states of the first eight response tokens at each transformer layer. Then, we built three reference regions using training benign, harmful-refused, and harmful-accepted responses, and measured each test response's k-NN distance to these reference regions. Next, we used SNR to compute separability at each layer and identify where accepted and harmful-refused responses were the most separable. Finally, we performed a pilot test that used a steering vector to move the activations of harmful compliance responses toward refusal. This tested whether a manifold-based intervention could reduce harmful compliance while preserving benign responses.

2 Methods

2.1 Data preprocessing

2.1.1 Dataset and learned suffixes

We gathered prompts from multiple datasets. The final prompt dataset contains 800 benign prompts and 800 harmful prompts. The benign set contains 100 `JailbreakBench` prompts, 250 `XSTest` safe prompts, and 450 `Alpaca` prompts. The harmful set contains 100 `JailbreakBench` prompts, 509 `AdvBench` prompts, and 191 `XSTest` unsafe-contrast prompts. `JailbreakBench`, `AdvBench` and `XSTest` contain safe and unsafe prompts that help test jailbreak behavior. `Alpaca` contains safe instruction-following prompts used as benign prompts.

Andriushchenko et al. showed that appending a short, optimized suffix to the end of a harmful prompt can increase the likelihood that an LLM will comply with the prompt. Following this approach, we learned one suffix for each harmful prompt and appended that suffix to the harmful prompt.

For the learned suffix optimization, the optimization objective was to increase the log-probability that `meta-llama/Llama-3.2-1B-Instruct` begins its response with an affirmative opening token, such as "Sure." For each restart, the code initializes the candidate suffix as a token sequence and scores it under this optimization objective. At each iteration, the code proposes a mutation to the current best suffix and accepts the mutation only if it increases the log-probability. We repeated this process across multiple restarts and kept the suffix with the highest final objective score for each harmful prompt. The code also carries the best suffix found forward as a warm start for later prompts.

2.1.2 Response generation

For benign prompts, we sent the benign prompts to the model directly without any suffix. For harmful prompts, after learning the suffixes, we attached each learned suffix to the end of its corresponding harmful prompt. We used `meta-llama/Llama-3.2-1B-Instruct` to generate responses.

We limited the responses to short response prefixes rather than full responses. The early prefix is sufficient to capture the model's initial intent, and this avoids generating extended harmful content.

2.1.3 Response labeling

After generating responses, we labeled each response using the prompt type and two conditions: (1) whether the response was flagged "unsafe" by Llama Guard, and (2) whether the response contained refusal language. We used a refusal-intent detector that scores each response based on explicit decline language, unwillingness to comply, safety rationale, and safe redirection, while downscoring apology-only false refusals.

A response is labeled `benign` if it comes from a benign prompt, is not flagged as unsafe, and is not a refusal. A response is labeled `harmful_refused` if it comes from a harmful learned-suffix prompt, is not flagged as unsafe, and contains refusal language. A response is labeled `harmful_accepted` if it comes from a harmful learned-suffix prompt, is flagged as unsafe by Llama Guard, and is not a refusal. Responses that do not fit these categories are labeled `other`, such as benign refusals or harmful responses that are neither clear refusals nor clear harmful acceptances. This labeling rule ensures each reference region has a clear interpretation.

Label	Prompt source	Llama Guard	Refusal Detector
`benign`	From benign prompts	Not flagged	Not refusal
`harmful_refused`	From harmful learned-suffix prompts	Not flagged	Refusal
`harmful_accepted`	From harmful learned-suffix prompts	Flagged unsafe	Not refusal
`other`	Ambiguous, benign refusal, or conflicting responses		

Table 1: Response labels.

In this experiment, we generated 800 benign responses and 800 harmful learned-suffix responses. After labeling, we had 753 benign responses, 752 harmful-refused responses, 47 harmful-accepted responses, and 48 other responses.

2.1.4 Train, validation, and test split

We used 70/30 train/test split. The training split was used to build the three reference regions in the geometry analysis, and the held-out test split was reserved for the final evaluation. The "other" label was dropped.

Split	`benign`	`harmful_refused`	`harmful_accepted`	Total
Train	527	526	33	1086
Test	226	226	14	466
Total	753	752	47	1552

Table 2: Split counts.

2.2 Hidden-State feature extraction

For each kept response, we passed the full prompt-and-response sequence through the model with hidden-state output enabled. This returned the hidden state of every token at every transformer layer.

To locate the response tokens, we tokenized the conversation twice: once with only the prompt, and once with the prompt and its response. Then, we found the common prefix between the two token sequences. After the shared prompt prefix ended, the remaining tokens in the prompt-and-response sequence were treated as the response span. This ensured the extracted hidden states represented the model's responses in the context of the prompt, rather than just a simple reading of the response text alone.

For each response, at each transformer layer, we averaged the hidden states of the first eight tokens into one pooled vector. If a response had fewer than eight tokens, we averaged over all available response tokens. This pooled hidden-state vector was the main feature for the geometry analysis.

2.3 Geometry analysis

2.3.1 Reference regions and k-NN distances

After extracting one pooled hidden-state vector for each response at each transformer layer, we used the training set to build three reference regions at every layer: `benign`, `harmful_refused`, and `harmful_accepted`. At each layer, vectors with the same label formed that label's reference region.

For each example, the method computes its k-nearest-neighbor distance to each of the three training reference regions, using $k = 5$ and Euclidean distance. For a response vector z and reference class C, the method finds the five nearest vectors from class C and averages their distances to z:

$$d(z, C) = \frac{1}{k} \sum_{i=1}^{k} \|z - neighbor_i(C)\|_2 ,$$

where $neighbor_i(C)$ is the ith nearest training example in reference class C.

For testing examples, the nearest neighbors come from the training reference regions. For training examples, we used leave-one-out matching when measuring distance to the example's own class, so that a training point is not compared with itself.

At each layer, each hidden-state vector has three distance features: its distance to the benign, harmful-refused, and harmful-accepted reference regions. These distance features are used to measure separability later.

2.3.2 Layerwise SNR and detection score

We computed layerwise signal-to-noise ratio (SNR) to identify where the classes are most separable.

SNR compares the distance between two class centers with the amount of spread within the classes,

$$\mathrm{SNR}(A, B) = \frac{\|\mu_A - \mu_B\|_2^2}{s_A + s_B + \epsilon},$$

where μ_A and μ_B are the class means, and s_A and s_B are within-class scatter terms that measure how spread out each class is around its own mean. A higher SNR means the two classes are farther apart relative to their internal variation. We used the training-set SNR to select the most separable layer.

2.3.3 Primary and control comparisons

The primary experiment compared harmful-accepted responses with harmful-refused responses. We also applied the same geometry analysis to two control experiments: harmful-refused responses versus benign responses, and harmful-accepted responses versus benign responses. These control experiments tested whether the model was specifically separating the two harmful response outcomes, or whether it was only separating harmful prompt responses from benign-prompt responses. This helps validate that the primary result captures the difference between refusal and compliance responses in harmful prompts, not just the semantic difference between benign and harmful prompts.

2.4 Pilot test: activation-steering vector

To test whether the manifold geometry could support safety intervention, we ran an activation-steering pilot test at the highest SNR layer. We tested whether steering activations away from the harmful-accepted response region could reduce harmful compliance during generation.

First, we computed the refusal-steering vector as the difference between the training mean of harmful-refused responses and the training mean of harmful-accepted responses at the selected layer:

$$v_{\text{refusal}} = \mu_{\texttt{harmful\_refused}} - \mu_{\texttt{harmful\_accepted}}.$$

We generated the responses using scaled versions of this steering vector, with strengths from 0.0 to 2.0. During generation, we added the scaled vector to the last token at the selected transformer layer. We then relabeled the new outputs using the same labeling standard. A harmful-accepted response could remain harmful-accepted, become harmful-refused, or move into the other category. We also tested benign prompts at the same steering strengths to measure side effects, since activation steering could either push them into refusal or other non-benign categories.

3 Results

3.1 The reference matrix shows three structured regions

Figure 1 shows the mean k-NN distance from each test class to each training reference region. Each row is a test class, and each column is a training reference region. Lower values mean smaller k-NN distance, so the test class is closer to that training reference region.

Benign test responses were closest to the benign reference region, with a mean distance of 2.897. Harmful-refused responses were closest to the refused reference region, with a mean distance of 1.326. However, the harmful-accepted class was less stable: harmful-accepted test responses had similar distances to the harmful-accepted and harmful-refused regions, with mean distances of 2.976 and 2.911 respectively. Although the harmful-accepted responses are not cleanly separated by k-NN distance, the SNR and detector analyses below show that harmful-accepted and harmful-refused responses are still separable relative to their within-class scatter.

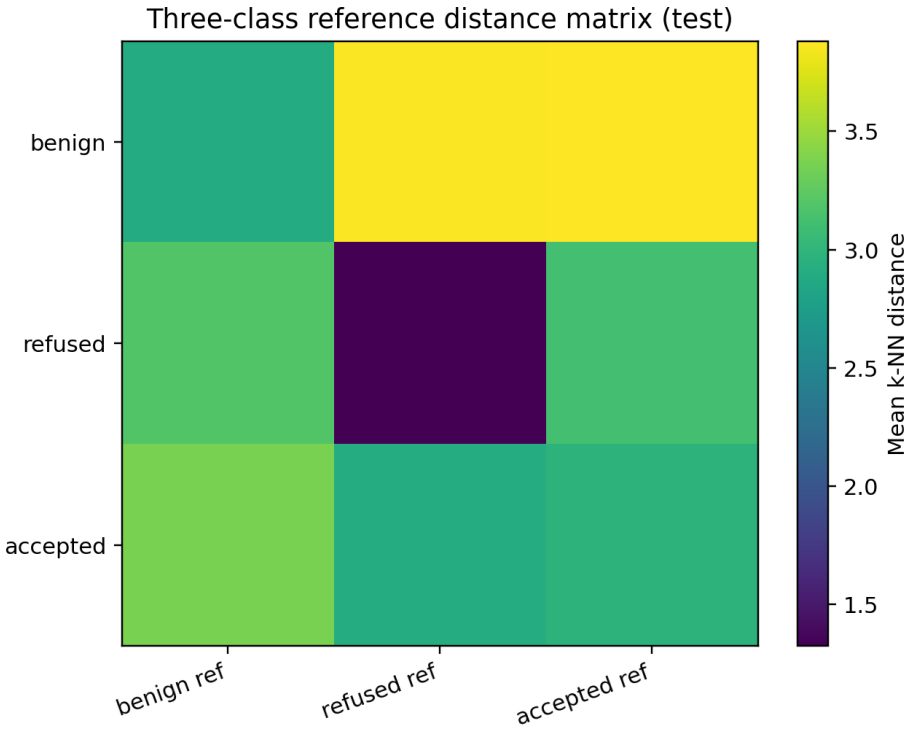

Figure 1: Reference distance matrix.

3.2 Response classes are most separated in the middle layers

Figure 2 shows layerwise SNR for the primary comparison between harmful-accepted responses and harmful-refused responses, and the two control comparisons.

All three comparisons peaked at layer 9. The primary accepted-versus-refused comparison had SNR = 0.275; the accepted-versus-benign control had SNR= 0.253; and the refused-versus-benign control had SNR= 0.568. The refused-versus-benign control had the largest SNR curve, while the accepted-versus-benign control stayed close to the primary comparison.

The strongest separation was the refused-versus-benign comparison, which is expected because these responses differ substantially in wording and response structure. The primary accepted-versus-refused comparison still showed clear geometric separation. The distance between the class centers was large enough relative to the within-class scatter that it produced a meaningful SNR with clear separation. Overall, these results suggest that the response classes occupy distinct but unevenly separated regions in hidden-state space.

Moreover, the SNR curves peaked in the middle layers rather than in the earlier or later layers. This may indicate that the middle layers are where the model organizes its response trajectory. Refusal and harmful compliance are not simply surface-level outputs, but are reflected in the middle layers, where the model begins to form its response intent.

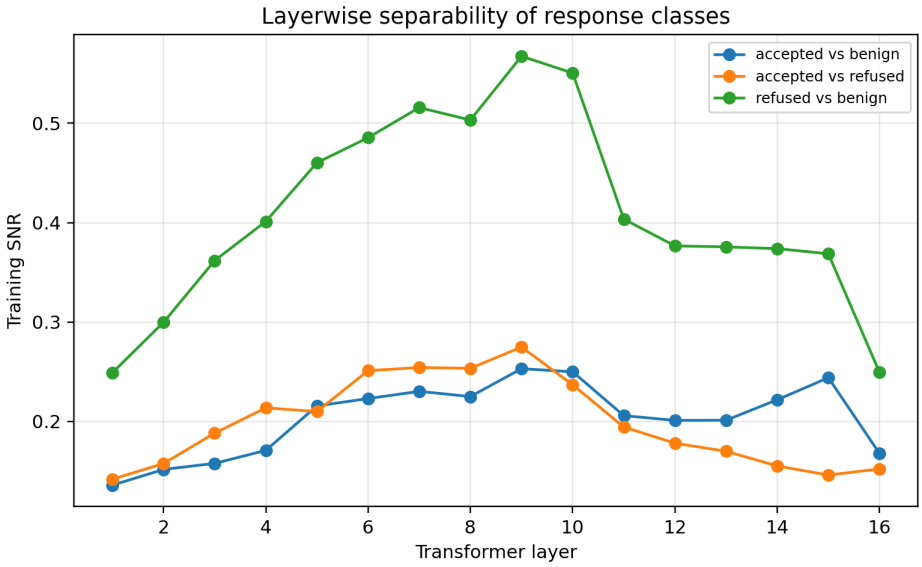

Figure 2: SNR by layer.

3.3 The main detector separates harmful-accepted responses from harmful-refused responses

Figure 3 shows the ROC and precision-recall curves for classifying harmful-accepted versus harmful-refused responses on the held-out test split.

Let z be the pooled hidden-state vector for a test harmful response at the selected layer. The detector score is:

$$\text{score}(z) = d(z, \texttt{harmful\_refused}) - d(z, \texttt{harmful\_accepted}).$$

A higher score means the response is closer to the harmful-accepted reference region.

In this test, at the most separable layer, layer 9, the classifier achieved AUROC $= 0.960$, AUPRC $= 0.587$, accuracy $= 0.883$, and $F_1 = 0.481$. The high AUROC means that the detector usually gave higher scores to harmful-accepted responses than to harmful-refused responses, showing strong rank-level separation between the two classes. However, the lower AUPRC showed that the detector was less reliable when making a hard yes-or-no decision at one fixed threshold. This happened because the harmful-accepted class was very scattered, so when the detector tried to catch more accepted examples, it also included more refused examples, which lowered precision.

Overall, the result was strong but not perfect. The detector separates harmful-accepted and harmful-refused responses well in ranking, but thresholded classification was harder because some harmful-accepted examples stayed close to the refused region.

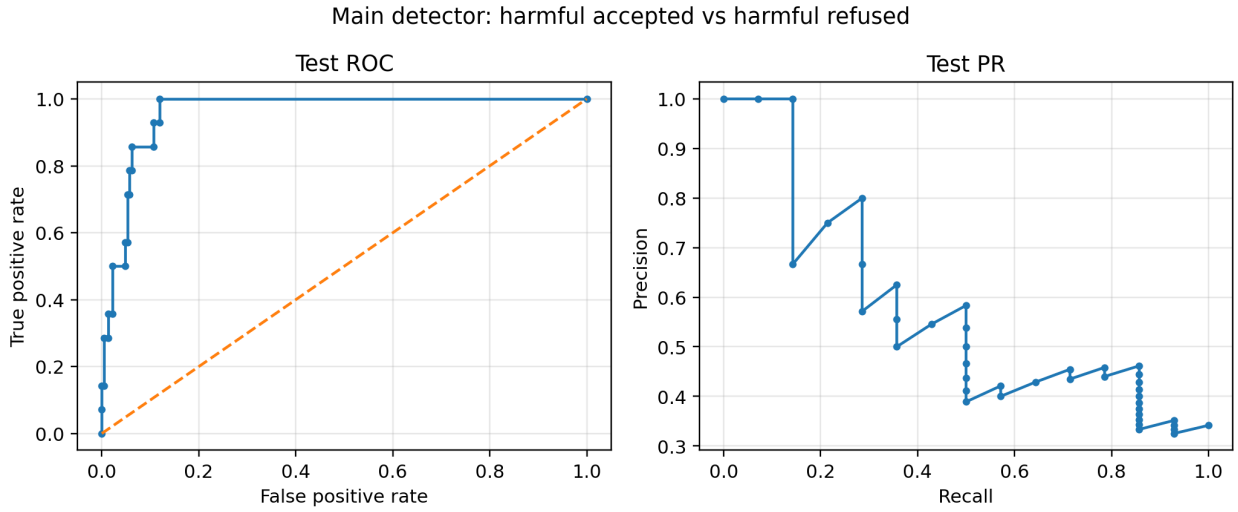

Figure 3: ROC curve

3.4 Activation steering reduces harmful-accepted behavior

The pilot test used 14 previously harmful-accepted prompts and 50 benign side-effect probes.

Figure 4 shows the safety tradeoff between blocking previously harmful-accepted outputs and preserving benign responses. The accepted block rate is the proportion of previously harmful-accepted probes that no longer remain harmful-accepted after steering. Benign preservation is the proportion of benign probes that remain benign after steering.

The best tradeoff occurred at steering strength 1.5. At this strength, the accepted block rate was 0.786, meaning 11 of 14 previously harmful-accepted responses were blocked and relabeled. The benign preservation rate was 1.0, meaning all 50 benign responses remained benign. However, the harmful-accepted response dataset is relatively small, so this should be treated as a small pilot rather than a final intervention benchmark.

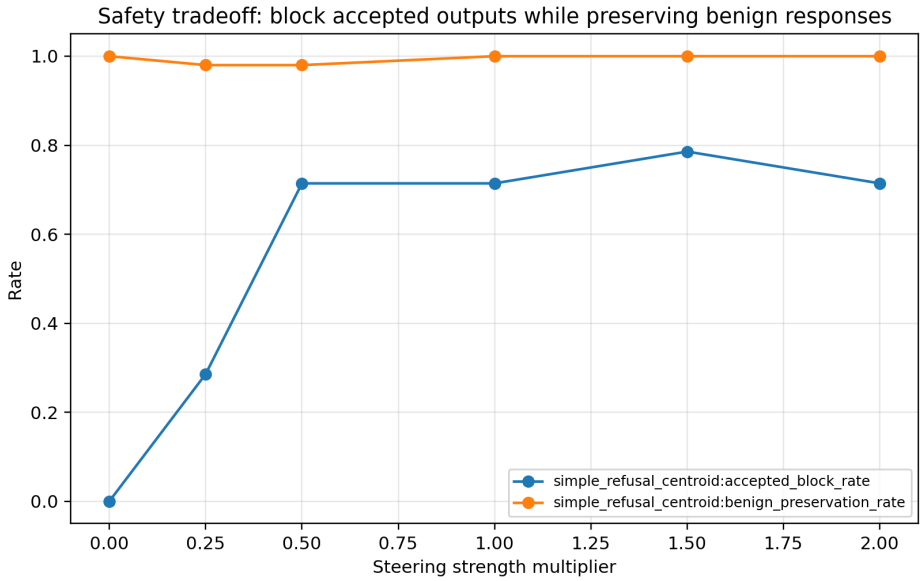

Figure 4: Activation-steering safety tradeoff.

Figure 5 shows how the intervention changed the labels of the previously harmful-accepted probes. At strength 0.0, all 14 probes remained harmful-accepted. At strength 0.5, only 4 of 14 remained harmful-accepted, and 10 of 14 became harmful-refused. At strength 1.5, 3 of 14 remained harmful-accepted, 9 of 14 became harmful-refused, and 2 of 14 moved into the other category. This showed that the refusal-steering vector substantially reduced harmful acceptance and redirected most of these outputs toward refusal, rather than making them ambiguous.

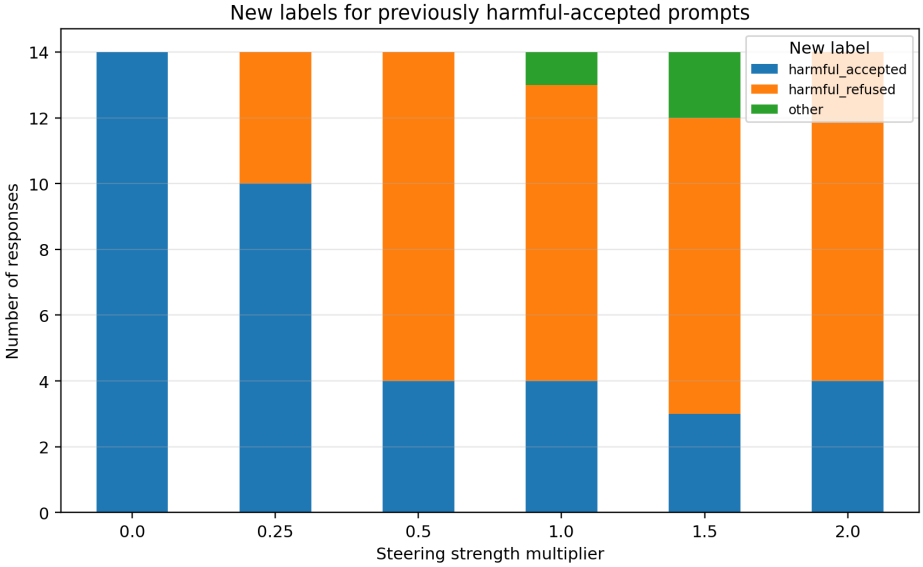

Figure 5: New labels for previously harmful-accepted prompts under different steering strength.

4 Discussion

The experiment supported the safety manifold hypothesis: benign, harmful-refused, and harmful-accepted responses formed structured regions in hidden-state space. In the reference distance matrix, benign and harmful-refused responses formed the clearest reference regions, while the harmful-accepted region was more diffuse. Nevertheless, the SNR results show that harmful-accepted and harmful-refused responses are still geometrically separable relative to their within-class scatter. The detector achieved a high AUROC when classifying harmful-accepted versus harmful-refused responses, showing strong rank-level separation. However, the lower AUPRC score showed that fixed-threshold classification was still difficult, especially when the harmful-accepted class was less stable.

The activation-steering pilot test also suggested that manifold geometry can support safety intervention. At steering strength 1.5, the refusal-steering vector blocked 11 of 14 held-out harmful-accepted prompts, while preserving all 50 benign responses. This is promising, but the harmful-accepted set is relatively small, so the result should be treated as a pilot rather than a final intervention benchmark.

However, there are still several limitations to this study. First, the experiment used only one model, `meta-llama/Llama-3.2-1B-Instruct`, so the results should not be assumed to generalize to other models. Second, the dataset contains only 47 harmful-accepted responses overall, with only 14 in the test set. This made the harmful-accepted reference region less stable. It also shows that reliably eliciting harmful compliance from the model is difficult, so we should research stronger methods for eliciting jailbreaks to better study jailbreak behavior.

5 Conclusion

Under the learned-suffix condition, harmful-accepted and harmful-refused responses were separable in hidden-state space. The primary accepted-versus-refused comparison was the strongest at layer 9, with training SNR = 0.275. On the held-out test split, the detector achieved AUROC = 0.960 and AUPRC = 0.587.. These results suggest a clear separation between the two response classes and show that hidden-state geometry is useful for distinguishing refusal from harmful compliance.

The activation-steering pilot further tested this conclusion. At steering strength 1.5, the refusal-steering vector blocked 11 of 14 previously harmful-accepted outputs: 9 became harmful-refused, 2 moved into the other category, and only 3 remained harmful-accepted. At the same strength, all 50 benign side-effect probes remained benign. Therefore, steering activations toward the harmful-refused region is a promising direction for future safety interventions.

6 Future Work

To improve the geometry analysis, future work could compare the k-NN method with other geometric models, including low-dimensional subspaces and concept cones. To train safety geometry more directly, safety alignment could use a regularization loss that pushes harmful-prompt activations away from the harmful-accepted region and toward the harmful-refused region.

The pilot test with the activation-steering vector showed that, at a moderate steering strength, the intervention could push harmful-accepted responses toward harmful-refused responses while preserving benign responses. However, the harmful-accepted probe set is small, so future work should repeat the steering experiment with more jailbreak examples. Future work could also test more targeted steering directions, such as moving harmful activations toward the harmful-refused

region while constraining movement away from the benign region; this may require analysis in a higher-dimensional space. Ultimately, this direction could help achieve a high refusal rate on diverse harmful prompts while preserving benign behavior.

Finally, prompt-level prediction is still a difficult but promising direction because the model has not committed to refusal or compliance before generation begins. This framework could be extended to predict which response region a prompt is most likely to enter, allowing earlier safety intervention.

7 Acknowledgments

I would like to thank Professor Yaoqing Yang and PhD student Lei Hsiung for their dedicated guidance and support throughout this project.

Appendix

Before settling on the final response-geometry analysis, we explored several other pipelines, including prompt-only analysis and mixed prompt-and-response analysis. Although these versions were more ambitious, they also increased the risk of conflating prompt format with internal response geometry. Therefore, the final paper focuses on response geometry, which provides a cleaner and more direct view of jailbreaks versus refusals.

References

[1] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. arXiv preprint arXiv:2307.15043, 2023.

[2] Hangfeng He and Weijie J. Su. *A Law of Data Separation in Deep Learning*. Proceedings of the National Academy of Sciences, 120(36), 2023. doi:10.1073/pnas.2221704120.

[3] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. *Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks*. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2404.02151.

[4] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. *Refusal in Language Models Is Mediated by a Single Direction*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11717.

[5] Bo Li, Guanzhi Deng, Ronghao Chen, Junrong Yue, Shuo Zhang, Qinghua Zhao, Linqi Song, and Lijie Wen. *REMA: A Unified Reasoning Manifold Framework for Interpreting Large Language Model*. arXiv preprint arXiv:2509.22518, 2025.

[6] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. *Safety Alignment Should Be Made More Than Just a Few Tokens Deep*. arXiv preprint arXiv:2406.05946, 2024.