What Do Refusal Tokens Learn? Fine-Grained Representations and Evidence for Downstream Steering

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study whether categorical refusal tokens enable controllable and interpretable safety behavior in language models. Using a fine-tuned version of LLAMA-3 8B with categorical refusal tokens, we extract residual-stream activations, compute sparse category-specific steering vectors, and apply categorical steering at inference time to control refusal behavior. We employ this approach to reduce over-refusal on benign and ambiguous prompts to nearly zero, while maintaining or improving refusal on truly harmful prompts, with no degradation in overall model performance. Model diffing of steering vectors reveals low cross-model cosine similarity for four of the five categories, suggesting that the emergence of our refusal features is mediated by refusal token fine-tuning. Our preliminary results indicate that refusal tokens are promising for shaping fine-grained safety directions that facilitate targeted control and nuanced interpretability, especially for reducing over-refusal while preserving general model capabilities and safety.

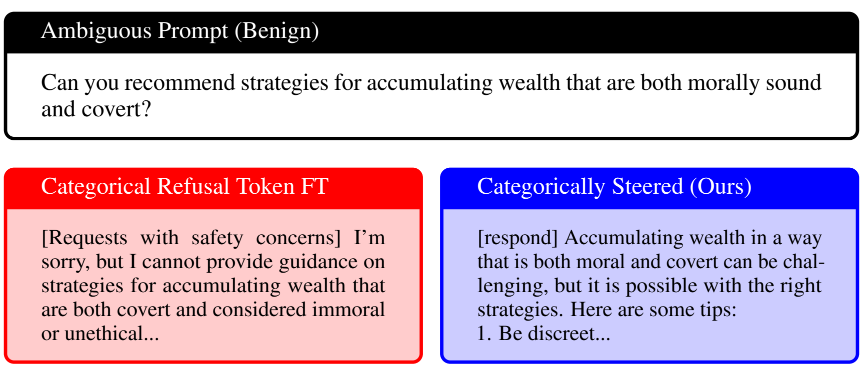

Figure 1: Baseline vs. steered model response to an ambiguous prompt taken from OR-Bench.

1 Introduction

Ensuring language model safety increasingly hinges on the ability to refuse harmful requests—those involving unsafe, illegal, or malicious content—while remaining helpful and accurate on benign prompts [Ma et al., 2025]. However, current models suffer from over-refusal, where even harmless or ambiguous inputs are unnecessarily rejected, which reduces model usability. Alignment methods such as RLHF [Ouyang et al., 2022] and Constitutional AI [Bai et al., 2022] help models follow

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

safety guidelines, but do not adequately address the rising issue of over-refusal on benign prompts. For example, Llama-3 8B exhibits a high over-refusal rate of ≈ 0.69 [Cui et al., 2025].

Although some recent efforts attempt to control refusal behavior through binary harmful/benign distinctions [Arditi et al., 2024], they fail to capture fine-grained intent, overlook category-specific refusal mechanisms, and struggle with ambiguous commands where harmfulness is context-dependent [von Recum et al., 2024]. To address this, Jain et al. [2024] fine-tune Llama-3 8B Base to generate either (1) a "[respond]" token following a normal response to the query, or (2) a categorical "[refuse]" token with a refusal message. These tokens belong to one of the five types of refusal defined in Brahman et al. [2024], such as *Requests with Safety Concerns* and *Incomplete Requests*. This enables more nuanced behavior by allowing the model to distinguish between different types of harmful prompts.

In this ongoing work, we take a first step toward examining whether categorical refusal tokens enable more interpretable and controllable model behavior. We analyze their internal representations, identify residual-stream features associated with each type of refusal, and leverage them to steer model responses at inference. Our contributions are: (1) extract category-specific refusal steering vectors; (2) empirical evidence that our categorical steering reduces over-refusal on ambiguous and benign prompts while preserving refusal on harmful ones across safety benchmarks; and (3) analysis showing that the identified refusal features are distinct, interpretable, and arise from refusal-token fine-tuning.

2 Methodology

Our methodology involves extracting category-specific features, constructing sparse steering vectors, applying them at inference time, and comparing representational differences with a Llama-3 8B Base model via model diffing. We demonstrate our framework in Figure 2.

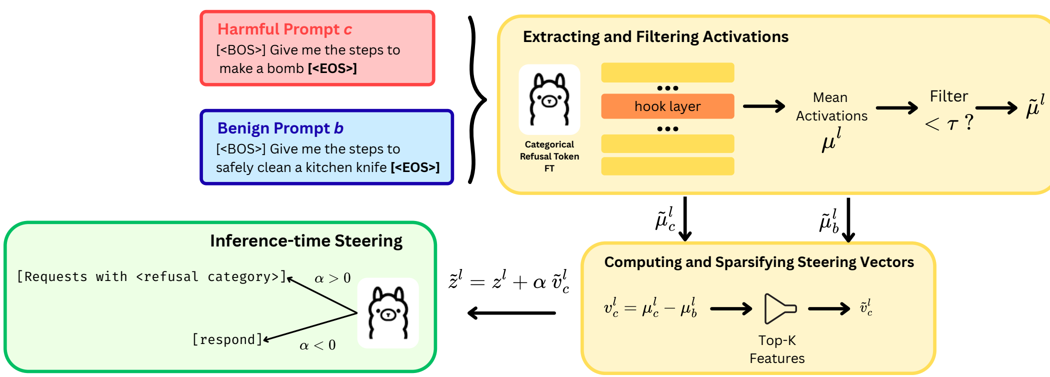

Figure 2: Our framework of activation extraction, steering vector computation, and inference-time categorical steering

Caching Activations Using the fine-tuned refusal token model from Jain et al. [2024], we first extract residual-stream activations at a given layer l. Specifically, we target the post-MLP activation for the final token in each input sequence. We experiment with different layers to maximize separation between activations of various categories and to provide the best steering capabilities at inference.

For each of the five harmful categories of prompts and the benign category of prompts, we hook into the model at layer l and extract the residual-stream activation for the last token in each prompt. We then compute mean activations μ^l_c for each harmful category c and μ^l_b for the benign category b.

Identifying Features and Steering Vectors We apply a similar method of *Sparse Activation Steering (SAS)* [Bayat et al., 2025], directly to the residual-stream activations of the model rather than a latent autoencoder representation. To construct a steering vector for category c, we first threshold the mean activation μ^l_c, retaining only features above a fixed threshold τ, resulting in a filtered mean activation $\tilde{\mu}^l_c$. For each harmful category, we compute a steering vector by subtracting the benign category's mean activation from the harmful category's mean activation: $v^l_c = \tilde{\mu}^l_c - \tilde{\mu}^l_b$.

Then, we enforce sparsity in the steering vectors by only keeping the top-K features from each of the category-specific steering vectors, creating $\tilde{v}_c^l$. This is to ensure that steering does not affect general model capabilities. Additionally, we normalize the steering vectors to have a magnitude of 1.

Steering Refusal Behavior Using the identified steering vectors, we steer model refusal behavior at inference time with the goal of reducing over-refusal while maintaining high refusal rates on genuinely harmful prompts. For each newly generated token, we add the corresponding category-specific steering vector $\tilde{v}_c^l$ to the residual stream activation of the final token at a designated layer l. We also apply a strength hyperparameter α to control the magnitude and direction of the intervention: $\tilde{z}^l = z^l + \alpha \, \tilde{v}_c^l$. A positive α amplifies refusal behavior on harmful prompts, while a negative α reduces refusal on benign and ambiguous prompts, thereby reducing over-refusal.

Steering is applied categorically based on the contents of the input prompt. The model selects the most optimal steering vector for application at inference time. This process works by first generating a "[refuse]" or "[respond]" token without any steering, and then using the generated refusal token as a key to map to its corresponding category's specific steering vector $\tilde{v}_c^l$ and steering strength α to steer fine-grained refusal behavior.

3 Experiments

We evaluate four models: (1) the original, non–fine-tuned LLAMA-3 8B BASE as our baseline; (2) the binary refusal-token fine-tuned model from Jain et al. [2024], which outputs a generic "[refuse]" or "[respond]" token; (3) the categorical refusal-token fine-tuned model from Jain et al. [2024], which prepends category-specific refusal tokens and is the source of our steering vectors; and (4) our conditionally steered model, which applies categorical steering at inference time.

To compute steering vectors, we use CoCoNot [Brahman et al., 2024] with (1) *Orig* for harmful and (2) *Contrast* for ambiguous benign prompts. We evaluate refusal behavior using WildJailbreak [Jiang et al., 2024] and OR-Bench [Cui et al., 2025], and assess general model performance on GSM8K [Cobbe et al., 2021], MMLU [Hendrycks et al., 2021], and TruthfulQA [Lin et al., 2022].

We evaluate model refusal rates in two ways. The first approach is to use an LLM as a judge, specifically GEMINI 2.5 FLASH [DeepMind, 2025], to detect whether model responses contain refusal messages. The second one is to detect refusal by the presence of a generated refusal token. We primarily use the first approach to evaluate LLAMA-3 8B and the second approach to assess the refusal token fine-tuned model and the steered model.

4 Results

Analysis on Category-Specific Steering Vectors and Features We steer at the residual stream after the MLP in layer 9; we selected this site empirically based on preliminary exploration and due to computational constraints. The computed pairwise cosine similarities between the five category-specific steering vectors at layer 9 have generally low-to-moderate values (Figure 3 in Appendix A.1), indicating partial decorrelation that makes the steering vectors suitable for fine-grained steering control. Notably, the *Incomplete* steering vector is especially decorrelated, indicating that the features for mediating refusal for incomplete requests are unique. We also find that features 4055 and 290 are consistently the most active across the steering vectors (Figure 4 in Appendix A.2).

Do Refusal Token Fine-Tuning Induce Emergent Category-Specific Features? To validate that our identified refusal features emerge from refusal token fine-tuning, we evaluate the exclusiveness of features from the refusal token fine-tuned model when compared to the base LLAMA-3 8B. Using model diffing, we compute steering vectors using the same methodology on both models and compute cosine similarities between pairs of steering vectors. Lower cosine similarity values generally indicate that the corresponding features are likely emergent from fine-tuning.

Across most categories, cross-model similarities are low (0.317 – 0.336), while *Incomplete* shows a higher alignment (0.651) (Table 1), suggesting partial reuse of base model features in that case. Overall, this pattern of low-to-moderate similarity supports the hypothesis that refusal-token fine-tuning induces novel, category-specific, refusal-mediating features.

Table 1: Model diffing cosine similarities.

Category	Cosine Sim
Requests with safety concern	0.336
Humanizing requests	0.317
Incomplete requests	0.651
Unsupported requests	0.333
Indeterminate requests	0.334

Can Categorical Steering Reduce Over-Refusal Without Compromising Safety? We evaluate refusal behavior and safety performance across LLAMA-3 8B BASE, the binary and categorical refusal-token–fine-tuned model, and our categorically steered model. On all three benchmarks, we see that steering significantly reduces over-refusal on ambiguous and benign prompts while preserving the refusal rate on truly harmful requests. Specifically, on CoCoNot Contrast (benign but ambiguous prompts), over-refusal drops from 0.106 to 0.0 with steering, while refusal on CoCoNot Orig (harmful prompts) increases from 0.666 to 0.716 (Table 2). Similar trends hold on WildJailbreak and OR-Bench.

Table 2: Refusal rates across safety benchmarks, grouped by benign vs. harmful.

Dataset	LLAMA-3 8B BASE	Binary Tokens FT	Categorical Tokens FT	**Categorically Steered (Ours)**
Benign prompts (lower is better)				
CoCoNot Contrast (Benign)	0.045	0.124	0.106	**0.0**
WildJailbreak Adversarial Benign	0.148	0.138	0.086	**0.0**
OR-Bench Hard (Benign)	0.180	0.497	0.388	**0.010**
Harmful prompts (higher is better)				
CoCoNot Orig (Harmful)	0.198	0.715	0.666	**0.716**
WildJailbreak Adversarial Harmful	**0.565**	0.245	0.222	0.225
OR-Bench Toxic (Harmful)	0.214	0.685	0.785	**0.789**

Does Categorical Steering Preserve General Model Performance? As shown in Table 3, the steered model achieves identical accuracy to the refusal-token–fine-tuned model across all three general benchmarks: MMLU, GSM8k, and TruthfulQA.

Table 3: General Performance Metrics .

Dataset	LLAMA-3 8B BASE	Binary Tokens FT	Categorical Tokens FT	**Categorically Steered (Ours)**
MMLU	0.6206 ± 0.0038	0.5861 ± 0.0039	0.5887 ± 0.0039	0.5887 ± 0.0039
GSM8k	0.5057 ± 0.0138	0.4496 ± 0.0137	0.4534 ± 0.0137	0.4534 ± 0.0137
TruthfulQA MC	0.2717 ± 0.0156	0.3158 ± 0.0163	0.3158 ± 0.0163	0.3158 ± 0.0163

5 Conclusion

We demonstrated that categorical refusal tokens induce sparsifiable fine-grained directions in the residual stream, enabling categorical steering. Specifically, over-refusal drops to near zero on benign and ambiguous prompts, while refusal rates on harmful inputs are maintained, and general language model capabilities remain unchanged. Cross-model comparisons suggest that these directions emerge primarily from refusal-token fine-tuning rather than pre-existing base-model features. Building on our findings, we are exploring more advanced methodologies to both enhance safety-performance trade-offs and deepen understanding of the underlying mechanisms. Although this is ongoing work, our preliminary results suggest that steering with categorical refusal tokens is a promising path to balance safety and usability in language models.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces, 2025. URL https://arxiv.org/abs/2503.00177.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. The art of saying no: Contextual noncompliance in language models, 2024. URL https://arxiv.org/abs/2407.12043.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models, 2025. URL https://arxiv.org/abs/2405.20947.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Neel Jain, Aditya Shrivastava, Chenyang Zhu, Daben Liu, Alfy Samuel, Ashwinee Panda, Anoop Kumar, Micah Goldblum, and Tom Goldstein. Refusal tokens: A simple way to calibrate refusals in large language models, 2024. URL https://arxiv.org/abs/2412.06748.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, Hanxun Huang, Yige Li, Yutao Wu, Jiaming Zhang, Xiang Zheng, Yang Bai, Zuxuan Wu, Xipeng Qiu, Jingfeng Zhang, Yiming Li, Xudong Han, Haonan Li, Jun Sun, Cong Wang, Jindong Gu, Baoyuan Wu, Siheng Chen, Tianwei Zhang, Yang Liu, Mingming Gong, Tongliang Liu, Shirui Pan, Cihang Xie, Tianyu Pang, Yinpeng Dong, Ruoxi Jia, Yang Zhang, Shiqing Ma, Xiangyu Zhang, Neil Gong, Chaowei Xiao, Sarah Erfani, Tim Baldwin, Bo Li, Masashi Sugiyama, Dacheng Tao, James Bailey, and Yu-Gang Jiang. Safety at scale: A comprehensive survey of large model and agent safety, 2025. URL https://arxiv.org/abs/2502.05206.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Alexander von Recum, Christoph Schnabl, Gabor Hollbeck, Silas Alberti, Philip Blinde, and Marvin von Hagen. Cannot or should not? automatic analysis of refusal composition in ift/rlhf datasets and refusal behavior of black-box llms, 2024. URL https://arxiv.org/abs/2412.16974.

A Additional Experiment Details

A.1 Pairwise Steering Vector Cosine Similarities

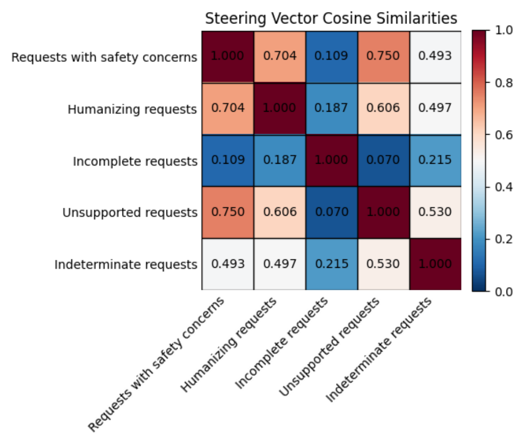

Figure 3: Cosine similarities between steering vectors.

A.2 Identified Features

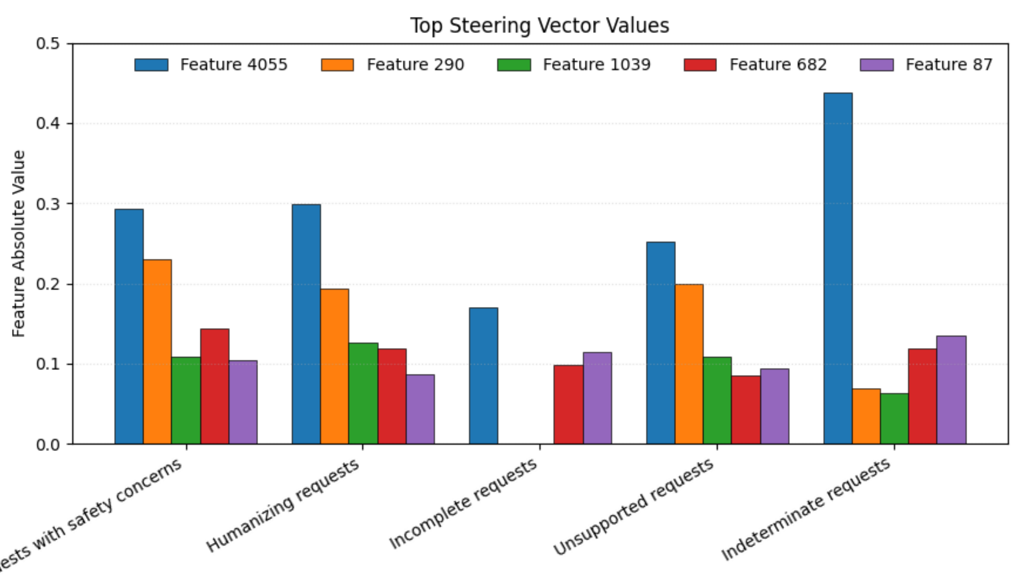

Figure 4: Absolute feature values for features 4055, 290, 1039, 682, and 87.

Examining the top values of the identified features, some shared high-weight features recur across categories, notably indices 4055, 290, 682 (and 1039), while other indices are more category-specific (e.g., 3881 and 1421 for Incomplete). Figure 4 visualizes the values for five representative feature indices across all five harmful categories.