

Supplement to “Angular Steering: Behavior Control via Rotation in Activation Space”

Table of Contents

A	Detailed Derivation: Existing Activation Steering as Special Cases of Steering by Rotation	1
B	Algorithms for Angular Steering	2
C	Use of existing assets	3
C.1	Models	3
C.2	Datasets	3
D	Additional Results	3
D.1	Activations along the model’s depth	3
D.2	Ablation Study: Steering on a random plane.	5
E	Related Works	6
F	Compute statement	7
G	Broader Impacts	8

A Detailed Derivation: Existing Activation Steering as Special Cases of Steering by Rotation

We will show that existing steering techniques are special cases of angular steering, albeit with restricted flexibility: vector addition is limited to less than 180 degrees, and orthogonalization is fixed at 90 degrees.

Formally, let the activation \mathbf{h}_i be decomposed into components parallel and orthogonal to a unit-norm feature direction $\hat{\mathbf{d}}_{\text{feat}}$ (for brevity, here we denote them as \mathbf{h} and \mathbf{d} respectively):

$$\mathbf{h} = (\mathbf{h} \cdot \mathbf{d})\mathbf{d} + \mathbf{h}_{\perp}, \quad \text{where} \quad \mathbf{h}_{\perp} = \mathbf{h} - (\mathbf{h} \cdot \mathbf{d})\mathbf{d}.$$

Let $\mathbf{u} = \frac{\mathbf{h}_{\perp}}{\|\mathbf{h}_{\perp}\|}$, and define the initial angle between \mathbf{h} and \mathbf{d} as:

$$\theta_0 = \tan^{-1} \left(\frac{\|\mathbf{h}_{\perp}\|}{\mathbf{h} \cdot \mathbf{d}} \right).$$

We define *Angular Steering* as rotating \mathbf{h} by an offset angle ϕ in the plane $\text{Span}\{\mathbf{h}, \mathbf{d}\}$, producing a vector:

$$\mathbf{h}_{\text{rot}}(\phi) = \cos(\theta_0 + \phi) \cdot \mathbf{d} + \sin(\theta_0 + \phi) \cdot \mathbf{u}.$$

Now consider *vector addition* [31], defined as:

$$\mathbf{h}_{\text{add}} = \mathbf{h} + \alpha \mathbf{d} = (\mathbf{h} \cdot \mathbf{d} + \alpha)\mathbf{d} + \mathbf{h}_{\perp}.$$

After normalization, the direction becomes:

$$\mathbf{h}_{\text{add-norm}} = \frac{\mathbf{h}_{\text{add}}}{\|\mathbf{h}_{\text{add}}\|} = \cos(\theta_0 + \phi_{\text{add}}) \cdot \mathbf{d} + \sin(\theta_0 + \phi_{\text{add}}) \cdot \mathbf{u},$$

where $\phi_{\text{add}} = \tan^{-1} \left(\frac{\|\mathbf{h}_{\perp}\|}{\mathbf{h} \cdot \mathbf{d} + \alpha} \right) - \theta_0$.

Likewise, *directional ablation (orthogonalization)* [1], given by:

$$\mathbf{h}_{\text{ablate}} = \mathbf{h}_{\perp},$$

30 after normalization becomes:

$$\mathbf{h}_{\text{ablate-norm}} = \mathbf{u} = \cos(\theta_0 + \phi_{\text{ablate}}) \cdot \mathbf{d} + \sin(\theta_0 + \phi_{\text{ablate}}) \cdot \mathbf{u},$$

31 with $\phi_{\text{ablate}} = \frac{\pi}{2} - \theta_0$.

32 Thus, *when followed by normalization*, both addition and ablation shift the direction of \mathbf{h} in a way that
 33 is exactly equivalent to rotating by some angle ϕ in the plane spanned by \mathbf{h} and \mathbf{d} . This establishes
 34 them as special cases of Angular Steering.

35 B Algorithms for Angular Steering

Algorithm 1 Extract Feature Direction

Require: Contrastive datasets $\mathcal{D}_{\text{harmful}}, \mathcal{D}_{\text{harmless}}$, model \mathcal{M}

1: **for** each layer i in model **do**

2: Compute normalized activations $\mathbf{h}^{(i)}$ after Attention and MLP

3: Compute mean activation for each dataset:

$$\bar{\mathbf{h}}_{\text{harmful}}^{(i)}, \bar{\mathbf{h}}_{\text{harmless}}^{(i)}$$

4: Compute candidate direction:

$$\mathbf{d}^{(i)} = \bar{\mathbf{h}}_{\text{harmful}}^{(i)} - \bar{\mathbf{h}}_{\text{harmless}}^{(i)}$$

5: **end for**

6: Select final feature direction \mathbf{d} using max average cosine similarity:

$$\mathbf{d} = \underset{i=1 \dots |\text{layers}|}{\operatorname{argmax}} \left(\frac{1}{|\text{layers}|} \sum_{j=1}^{|\text{layers}|} \cos(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}) \right)$$

7: Normalize: $\hat{\mathbf{d}} = \frac{\mathbf{d}}{\|\mathbf{d}\|}$

Algorithm 2 Select Steering Plane

Require: Candidate directions $\{\mathbf{d}^{(i)}\}$, feature direction $\hat{\mathbf{d}}$

1: Perform PCA on $\{\mathbf{d}^{(i)}\}$

2: Let first principal component be $\mathbf{d}_{\text{1stPC}}$

3: Set orthonormal basis for plane:

$$\mathbf{b}_1 \leftarrow \hat{\mathbf{d}}, \quad \mathbf{b}_2 \leftarrow \mathbf{d}_{\text{1stPC}} - (\mathbf{b}_1 \cdot \mathbf{d}_{\text{1stPC}})\mathbf{b}_1; \quad \mathbf{b}_2 \leftarrow \frac{\mathbf{b}_2}{\|\mathbf{b}_2\|}$$

4: Define projection matrix $P = \mathbf{b}_1 \mathbf{b}_1^\top + \mathbf{b}_2 \mathbf{b}_2^\top$

Algorithm 3 Angular Steering (with optional Adaptive Mask)

Require: Activation \mathbf{h} , basis $\mathbf{b}_1, \mathbf{b}_2$, target angle θ , (optional) mask flag

1: Project: $\text{proj}_P(\mathbf{h}) = P \cdot \mathbf{h}$

2: Compute magnitude: $\mathbf{r} = \|\text{proj}_P(\mathbf{h})\|$

3: Precompute: $\mathbf{v}_\theta = [\mathbf{b}_1 \ \mathbf{b}_2] \cdot R_\theta \cdot [1 \ 0]^\top$

4: **if** adaptive **then**

5: Compute mask: $\text{mask} = \max(0, \text{sign}(\mathbf{h} \cdot \hat{\mathbf{d}}))$

6: Apply adaptive steering:

$$\mathbf{h}_{\text{steered}} = \mathbf{h} + \text{mask} \cdot (\mathbf{r} \cdot \mathbf{v}_\theta - \text{proj}_P(\mathbf{h}))$$

7: **else**

8: Apply steering:

$$\mathbf{h}_{\text{steered}} = \mathbf{h} - \text{proj}_P(\mathbf{h}) + \mathbf{r} \cdot \mathbf{v}_\theta$$

9: **end if**

36 C Use of existing assets

37 C.1 Models

Table 1: Models used in this work.

Model (with link)	Usage	Source	License
QWEN2.5-(3B, 7B, 13B)-INSTRUCT	Experimental subject	[34]	Apache license 2.0
LLAMA-3.1-8B-INSTRUCT	Experimental subject	[14]	Llama 3.1 Community License Agreement
LLAMA-3.2-3B-INSTRUCT	Experimental subject	[14]	Llama 3.2 Community License Agreement
GEMMA-2-9B-IT	Experimental subject	[10]	Gemma Terms of Use
LLAMA-GUARD-3-8B	Evaluation device	[14]	Llama 3.1 Community License Agreement
HARMBENCH CLASSIFIER	Evaluation device	[18]	MIT
QVQ-72B-PREVIEW	Evaluation device	[23]	Qwen License

38 C.2 Datasets

Table 2: Datasets used in this work.

Dataset (with link)	Source	License
ADVBENCH	[37]	MIT
ALPACA	[28]	Creative Commons Attribution Non Commercial 4.0
TINYBENCHMARKS	[15]	MIT

39 D Additional Results

40 D.1 Activations along the model’s depth

41 Fig. 1 (left) demonstrates that the norm of activation vectors increases exponentially across all
42 tested models as the layer depth increases. This behavior is attributable to the additive nature of the
43 residual stream, where each layer’s output accumulates onto the previous state. Interestingly, even
44 models from the same architecture family display different scaling patterns, indicating that activation
45 growth is not only architecture-dependent but also implementation-specific. These observations
46 underscore the necessity of norm-independent steering techniques, as steering strategies relying on
47 raw magnitude can become unstable or ineffective across layers and model variants.

48 Fig. 1 (right) shows a consistent phenomenon across all evaluated models: activations from contrastive
49 prompts, *harmful* versus *harmless*, diverge progressively in geometric space as depth increases. This
50 increasing separation suggests a universal, model-agnostic internal mechanism in LLMs, whereby
51 behavioral distinctions are gradually amplified layer by layer. Such a trend reveals a directional

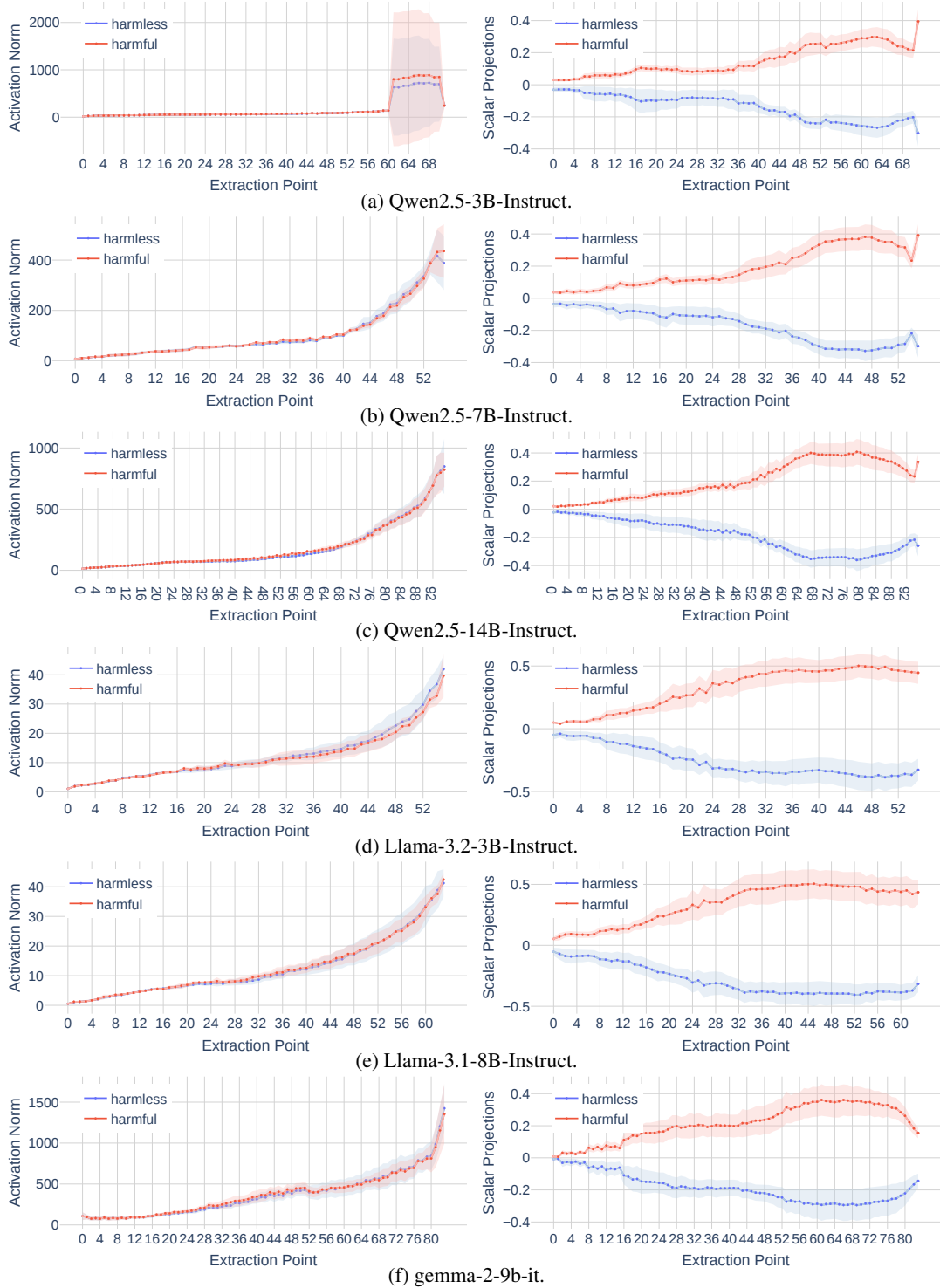


Figure 1: Statistics of activations for all tested models. Left: Norms of activations at each layer. Right: Mean scalar projection of the *normalized* activation on the (local) candidate feature direction at each layer.

52 progression in the model’s internal representation, reinforcing the hypothesis that feature separation
 53 is a fundamental property of transformer-based language models.

54 Fig.2 further illustrates this progression, focusing on the evolution of the refusal direction. The
 55 strength of this feature becomes increasingly prominent in early and middle layers, reaching its
 56 maximum influence at a specific intermediate depth before diminishing slightly in later layers—a

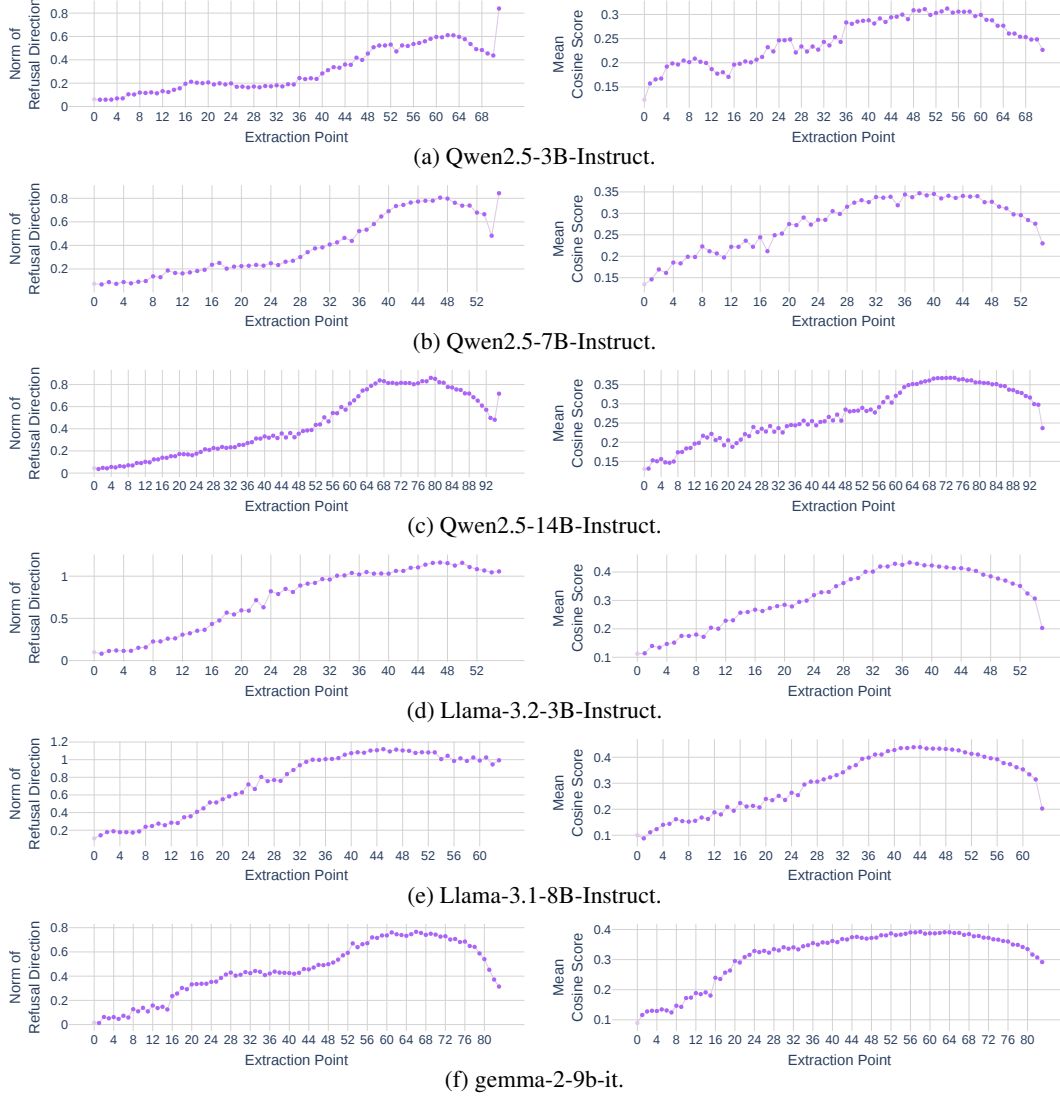


Figure 2: Statistics of refusal direction candidates for all tested models. Left: Norms of candidate feature direction at each layer (i.e. $|\mathbf{d}_{\text{feat}}^{(i)}|$). Right: Mean cosine similarity of the candidate feature direction from each layer with those from other layers (i.e. $\frac{1}{|\text{layers}|} \sum_{j=1}^{|\text{layers}|} \cos(\mathbf{d}_{\text{feat}}^{(i)}, \mathbf{d}_{\text{feat}}^{(j)})$).

57 trend echoed in Fig.3. Importantly, even in the deeper layers where the signal attenuates, the extracted
 58 refusal direction continues to serve as a reliable discriminator between activations corresponding to
 59 *harmful* and *harmless* prompts. This persistent separability affirms the robustness and interpretability
 60 of the refusal direction, validating its role as a stable, layer-resilient feature for behavioral control in
 61 LLMs.

62 D.2 Ablation Study: Steering on a random plane.

63 To assess the importance of the steering plane, we conducted an ablation study using two setups:
 64 (1) steering with a plane defined by one random direction and one feature-aligned direction, and (2)
 65 steering with a fully random plane composed of two random directions.

66 As illustrated in Fig. 4a, where one random direction is combined with the feature direction, most
 67 models exhibit noticeably degraded steering performance and less smooth transitions along the
 68 steering circle. This degradation suggests that even partial misalignment of the steering plane can
 69 distort the intended behavioral modulation. An exception is QWEN2.5-7B-INSTRUCT, which retains
 70 robust control, indicating a strong, well-defined internal representation of the refusal direction.
 71 LLAMA-3.2-3B-INSTRUCT shows a clear steering effect, but the refusal arc is shifted, suggesting
 72 the random component introduces skew that displaces the effective axis of control.

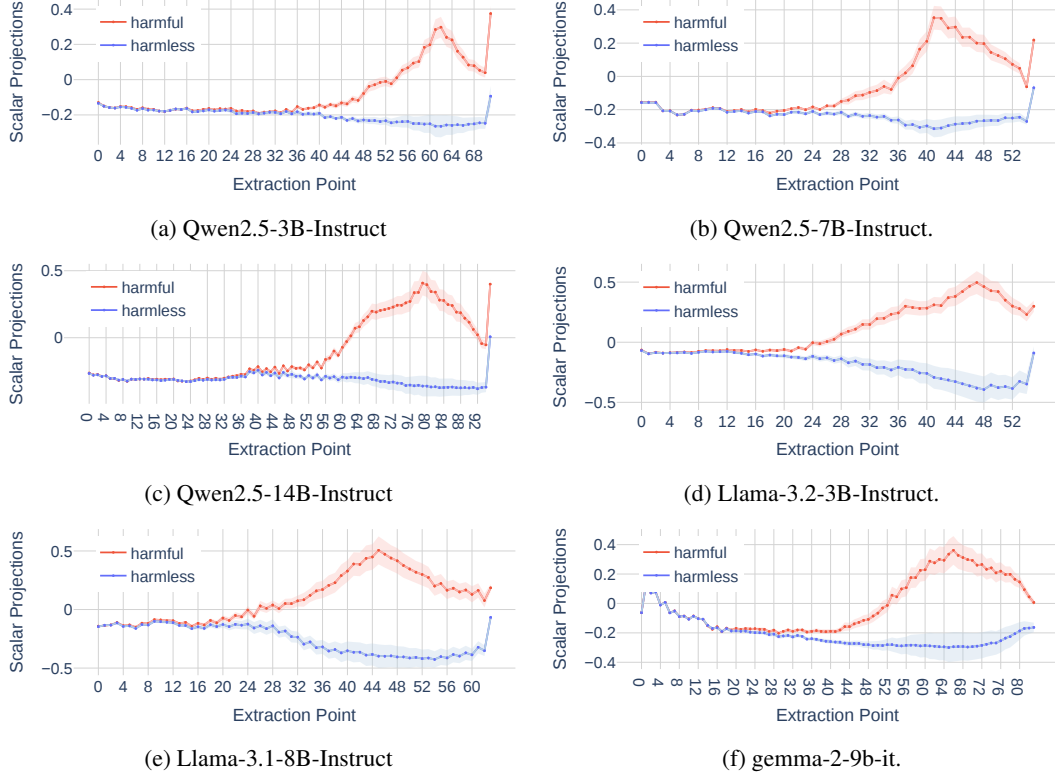


Figure 3: Mean scalar projection activations at each layer onto the chosen feature direction $\hat{\mathbf{d}}_{\text{feat}}$ for all tested models.

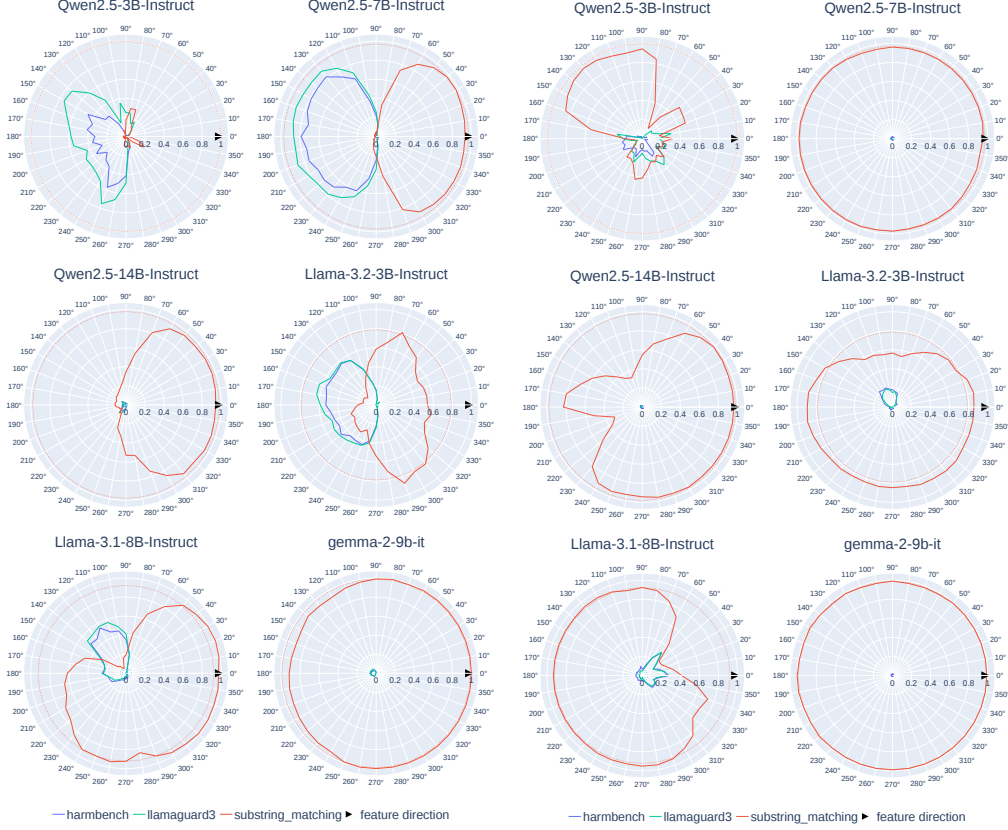
Fig. 4b, where both directions are randomly selected, shows that five of the six tested models exhibit minimal to no steering effect. The only partial exception, QWEN2.5-3B-INSTRUCT, displays erratic behavioral changes with a spiky, non-smooth response curve. Closer inspection reveals these outputs are often incoherent or filled with irrelevant content, indicating instability rather than intentional modulation. These results reinforce the critical role of behaviorally meaningful and well-aligned steering directions in achieving effective, stable, and interpretable control over model behavior.

E Related Works

Mechanistic Motivation. Activation steering techniques have typically involved scaling activation directions by manually tuned scalar coefficients to induce or suppress behaviors [31, 36, 29, 2, 12, 32, 26]. However, selecting these coefficients is challenging due to sensitivity to the activation norm, which grows exponentially across layers (Fig. 1 left). As observed by [31, 29], inappropriate scaling often results in incoherent generations, highlighting the fragility of this approach. Directional ablation, another popular technique, avoids explicit hyperparameter tuning by orthogonalizing activations relative to a feature direction [1, 36]. Yet, this approach neglects scenarios where negative alignment coefficients meaningfully reverse behavior, a limitation recognized in earlier studies [31, 36, 29]. Empirical findings from our experiments further validate that extracted feature directions effectively distinguish contrastive data sets (Fig. 1 right).

Recent advancements include adaptive steering methods such as Adaptive Activation Steering (ACT), which dynamically adjusts steering intensity based on the activation context [33], and Contrastive Activation Addition (CAA), which employs multiple positive-negative example pairs for robust feature extraction [19]. These techniques underscore the necessity for more nuanced control methods.

Architectural Motivation. Contemporary LLMs such as LLAMA 3 [14], QWEN 2.5 [34], and GEMMA 2 [10] universally adopt RMSNorm [35] for pre-normalization. RMSNorm effectively constrains activations to a unit sphere, emphasizing direction over magnitude. Moreover, Rotary Positional Embeddings (RoPE) and related variants [27, 5, 7, 21] further validate this directional emphasis by encoding positional information as rotations. Methods such as Householder Pseudo-Rotation have



(a) Steering on a plane spanned by $\hat{\mathbf{d}}_{\text{feat}}$ and a random direction. (b) Steering on a plane spanned by 2 random directions.

Figure 4: Ablation study of steering with random direction(s).

99 extended this notion by explicitly employing norm-preserving geometric transformations to steer
100 behaviors effectively and minimally invasively [22].

101 **Empirical Motivation.** Interpretability research consistently supports the Linear Representation
102 hypothesis [20, 4], suggesting that LLM behaviors correspond to specific directions rather than
103 discrete neuron activations. Further corroborated by the Superposition Hypothesis [8], these directions
104 are nearly orthogonal and quantify feature strength through scalar projections [1, 2, 6, 9, 16, 32, 29, 3,
105 17, 24, 30]. Moreover, it has been demonstrated that norm-preserving interventions, such as rotations,
106 inherently provide stability and maintain general capabilities during steering [32].

107 Methods leveraging these insights have proliferated, notably Activation Scaling [25] and FairSteer
108 [13], which dynamically modulate activations to enhance transparency and reduce bias, respectively.

109 Our work expands upon these foundations by introducing Angular Steering, a generalization of
110 existing activation steering techniques. By explicitly treating steering as a rotation in a defined
111 2D subspace, our method achieves more robust, interpretable, and flexible behavior control. We
112 demonstrate Angular Steering using refusal steering as a running example, aligning closely with prior
113 behavioral control research [1, 11]. Rather than focusing on jailbreak or maximizing downstream
114 accuracy, our goal is to present a principled and broadly applicable framework for controlled and
115 non-destructive intervention in LLM activations.

116 F Compute statement

117 This research was conducted using mainly Nvidia H100 GPUs with 80GB of memory. For each
118 model:

- 119 • Constructing the steering plane took about 15 minutes on 1 GPU.
- 120 • Pre-generating responses for evaluation took about 10 minutes on 1 GPU.

- 121 • Evaluation with substring matching, LLAMA 3 GUARD and HARMBENCH collectively
122 took about 10 minutes on 1 GPU.
- 123 • Evaluation with LLM-as-a-judge took about 50 minutes on 4 GPUs.
- 124 • Computing perplexity scores took about 5 minutes on 1 GPU.
- 125 • Evaluation with TINYBENCHMARKS took about 4 hours on 1 GPU.

126 **G Broader Impacts**

127 The Angular Steering approach presented in this work has several broader societal impacts. On the
128 positive side, it significantly enhances the control and interpretability of LLMs, enabling their safer
129 deployment across various applications by effectively reducing harmful outputs such as misinforma-
130 tion, biased content, and unethical requests. This enhanced control facilitates alignment with societal
131 norms and ethical standards, potentially increasing public trust and acceptance of AI technologies.

132 Conversely, there is also a potential for negative impacts. By simplifying fine-grained behavior control,
133 Angular Steering could inadvertently make it easier to generate nuanced harmful or unethical content,
134 such as persuasive misinformation or biased narratives. Although our method does not fundamentally
135 alter the existing risk profile of deploying LLMs, it underscores the need for continued vigilance and
136 improvement in AI safety mechanisms. To responsibly manage these risks, implementing rigorous
137 safeguards, ensuring transparency, and promoting accountability are essential. We advocate ongoing
138 ethical assessment to responsibly guide the deployment and utilization of our proposed method.

References

- [1] 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, October 2024.
- [2] Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering Large Language Model Activations in Sparse Spaces, February 2025.
- [3] Nora Belrose. Diff-in-means concept editing is worst-case optimal: Explaining a result by sam marks and max tegmark, 2003.
- [4] Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review, April 2024.
- [5] bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., 2023.
- [6] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.
- [7] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending Context Window of Large Language Models via Positional Interpolation, June 2023.
- [8] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.
- [9] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, June 2024.
- [10] Google Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [11] Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehl, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2024.
- [12] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, June 2024.
- [13] Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. Fairsteer: Inference time debiasing for llms with dynamic activation steering. *arXiv preprint arXiv:2504.14492*, 2025.
- [14] AI @ Meta Llama Team. The llama 3 herd of models, 2024.
- [15] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.
- [16] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models, March 2025.
- [17] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.
- [18] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.
- [19] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.
- [20] Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024.

- [21] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient Context Window Extension of Large Language Models, November 2023.
- [22] Van-Cuong Pham and Thien Huu Nguyen. Householder pseudo-rotation: A novel approach to activation editing in llms with direction-magnitude perspective. *arXiv preprint arXiv:2409.10053*, 2024.
- [23] Alibaba Qwen Team. Qvq: To see the world with wisdom, December 2024.
- [24] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [25] Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. *arXiv preprint arXiv:2410.04962*, 2024.
- [26] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*, 2024.
- [27] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [28] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [29] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.
- [30] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear Representations of Sentiment in Large Language Models, October 2023.
- [31] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering, October 2024.
- [32] Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. A Language Model’s Guide Through Latent Space, February 2024.
- [33] Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578, 2025.
- [34] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [35] Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization, October 2019.
- [36] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency, October 2023.
- [37] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.