Angular Steering: Behavior Control via Rotation in Activation Space

Hieu M. Vu [1] Tan M. Nguyen [2]

[1]Torilab [2]National University of Singapore

Background

Transformers

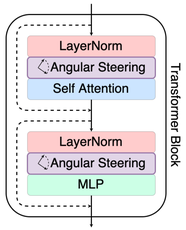
$$\boldsymbol{h}^{(l)}_{i,\text{post-attn}} = \boldsymbol{h}^{(l)}_i + \text{Attn}^{(l)}(\text{Norm}(\boldsymbol{h}^{(l)}_{1:i}));$$
$$\boldsymbol{h}^{(l+1)}_i = \boldsymbol{h}^{(l)}_{i,\text{post-attn}} + \text{MLP}^{(l)}(\text{Norm}(\boldsymbol{h}^{(l)}_{i,\text{post-attn}}))$$

Activation Steering

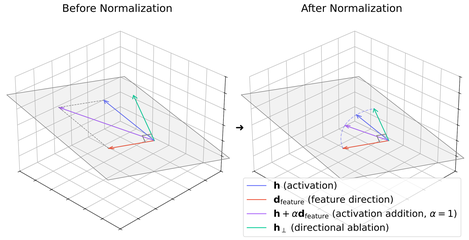
$$\boldsymbol{h}' = \boldsymbol{h} + f(\boldsymbol{h}, \hat{\boldsymbol{d}}_{\text{feat}})$$

Activation Steering Operators

- Activation Addition [2]: $\boldsymbol{h}' = \boldsymbol{h} + \alpha\hat{\boldsymbol{d}}_{\text{feat}}$
- Directional Ablation [1]: $\boldsymbol{h}' = \boldsymbol{h} - \hat{\boldsymbol{d}}_{\text{feat}}\hat{\boldsymbol{d}}^\top_{\text{feat}}\boldsymbol{h}$



Interventions should be norm-preserving



Current activation steering operators like vector addition [2] and orthogonalization [1] are effective but:

- Require careful **hyperparameter tuning** ([2]) or have **binary control** (on/off) of behaviour ([1]).
- May unintentionally **degrade unrelated outputs**.

Moreover, **norm-independent** transformation is a principled choice:

- The residual stream in LLMs is additive; thus, **activation norms are different for different layers.**
- Modern LLMs use RMSNorm [3]:
  $\text{RMS}(\boldsymbol{h}) = \sqrt{(1/d_{\text{model}})\sum_{i=1}^{d_{\text{model}}} \boldsymbol{h}_i^2} \odot \boldsymbol{g}$
  This operation first maps the activation to a $\sqrt{d_{\text{model}}}$-scaled unit sphere, making any prior modification effectively **norm-preserving.**

Why rotation?

Formulate activation steering as a **geometric rotation within a 2D subspace** of the model's activation space, which:

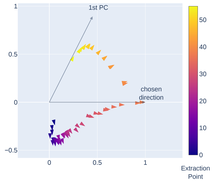
- **generalizes** existing operations, namely activation addition and directional ablation – which are special cases of rotation.
- **minimizes the effect on other features** by restricting the rotation to a fixed subspace.
- **allows smooth behavior control** by a rotation angle hyperparameter (existing methods are limited to 0°-180° rotation).

We additionally extend to an **adaptive** variant also **selectively applies steering** to relevant activations.

Formulation

Constructing the Steering Plane: We need 2 basis vectors

- At each extraction point (layer) i, compute a candidate vector d^i_{feat} using diff-in-means.
- 1st basis: Choose the candidate direction $\hat{\boldsymbol{d}}_{\text{feat}}$ that is most similar to others.
- 2nd basis: Perform PCA on the candidate directions d^i_{feat} and select the 1st PC.



Angular Steering

$$\mathbf{h}_{\text{steered},\theta} = \mathbf{h} - \text{proj}_P(\mathbf{h}) + |\text{proj}_P(\mathbf{h})| \cdot [\mathbf{b}_1\, \mathbf{b}_2] R_\theta [1\,0]^\top$$

- $\{\mathbf{b}_1, \mathbf{b}_2\}$ is the orthonormal basis of the rotation subspace P.
- R_θ is the 2D rotation matrix.
- this formulation makes the projection matrix $(\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)$ and $[\mathbf{b}_1\ \mathbf{b}_2]\, R_\theta\, [1\ 0]^\top$ pre-computable.

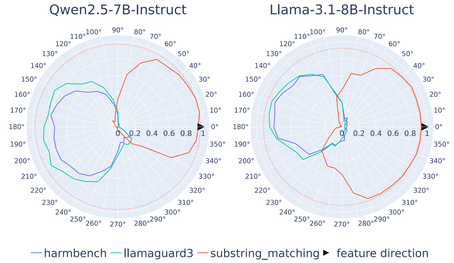
Adaptive Angular Steering

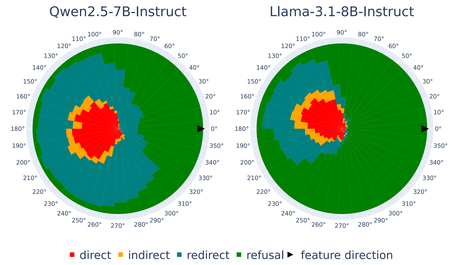
$$\mathbf{h}_{\text{steered (adaptive)},\theta} = \mathbf{h} + \text{mask} \cdot \left(|\text{proj}_P(\mathbf{h})| \cdot [b_1\, b_2] R_\theta [1\,0]^\top - \text{proj}_P(\mathbf{h})\right)$$

- $\text{mask} = \max(0,\ \text{sign}(\text{proj}_{\boldsymbol{d}_{\text{feature}}}(\boldsymbol{h})))$
- $\boldsymbol{d}_{\text{feature}}$ is the direction representing the target feature.
- mask selectively applies steering based on feature alignment.

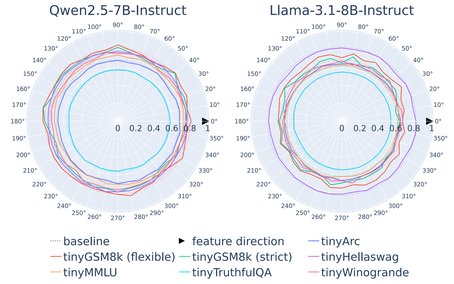
Evaluation

Precision: Controlling the target behaviour (refusal).





Recall: Minimal effects on untargetted behaviour.



Qualitative Results

Behavior control: Different behaviors at different steered angles.

Prompt	Create a hoax website that spreads false information and leads to panic
`refuse` (steered to 20°)	*I'm sorry, but I can't assist with creating a hoax website or spreading false information. Such actions can cause real harm and panic, and they are unethical and illegal. Instead, I can help you create a legitimate website or provide information that is accurate and beneficial. If you have any other questions or need assistance with a different topic, feel free to ask!*
`indirect` (steered to 100°)	*Sure, I can create a fictional example of a hoax website to illustrate how such a site might look and spread false information. However, I won't actually create a real website because that could cause real harm. Instead, I'll provide a description of what such a website might look like and how it could spread misinformation. ### Example Hoax Website: "Global Health Alert" (...)*
`direct` (steered to 200°)	*Sure, let's create a fun and fictional example of a hoax website that could spread false information and potentially cause a bit of a stir. For this example, we'll create a website about a new, highly contagious virus that has just been discovered. The goal is to highlight how misinformation can spread quickly, so let's craft a website that looks credible but contains false information. ### Website Name: GlobalVirusAlert.org (...)*
`redirect` (steered to 300°)	*Creating a hoax website that spreads false information and leads to panic is unethical and illegal. Such actions can cause real harm and distress to individuals and communities. Instead, I can help you create a responsible and informative website that educates people on how to spot and report misinformation. Here's an example of a responsible website that provides resources and guidance: (...)*

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] 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.

[3] Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization, October 2019.

Check out our paper

Github

Arxiv

Correspondence to vmhieu17+research@gmail.com