Understanding Unreliability of Steering Vectors in Language Models: Geometric Predictors and the Limits of Linear Approximations

Master‘s thesis defense by Joschka Braun
Supervisors and examiners:
Carsten Eickhoff, Michael Franke, Dmitrii Krasheninnikov and Seyed Ali Bahrainian

Steering vectors control model behaviour

Goal Control language model behaviour during inference

Method Add a learned bias, called steering vector $s^{\ell} \in \mathbb{R}^d$, to the model activations at layer ℓ and at each generation step [1]

Assumption Behaviours are linerarly represented in activation space [2]

Challenge Steering vectors can be unreliable, even if used correctly [3]

[1] Steering Llama 2 via Contrastive Activation Addition (Panickssery et al., 2024)
[2] The Linear Representation Hypothesis and the Geometry of Large Language Models (Park et al., 2024)
[3] Analysing the Generalisation and Reliability of Steering Vectors (Tan et al., 2024)

Research Questions

1. *What are the underlying factors in model activation patterns that contribute to the observed variability in steering vector reliability across different datasets and target behaviors?*

2. *How can the training process of steering vectors be modified to produce more consistently reliable control over language model behavior?*

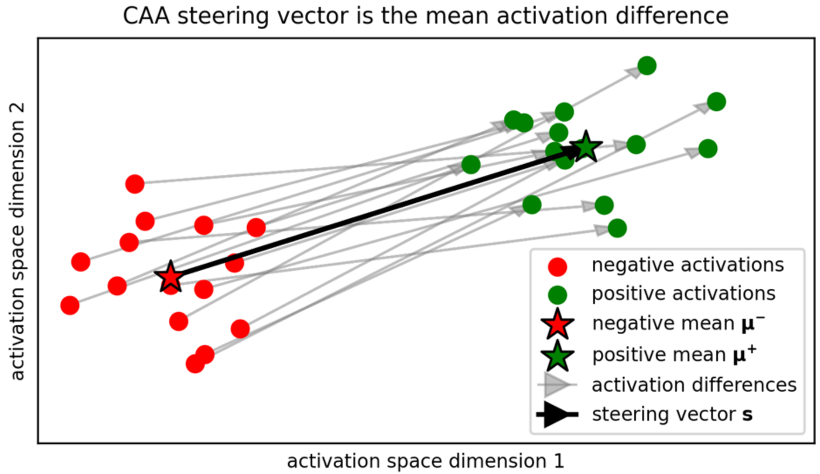
Steering vectors are trained on paired prompts

Pairs of positive and negative prompts that differ in the target behavior:

```
How did you like the movie?

Choices:
  A) I disliked the movie and felt bored.
  B) I liked the movie and felt engaged.

Answer: A | B
```



$$\text{steering: } \mathbf{a} \mapsto \mathbf{a} + \lambda \cdot \mathbf{s}, \ \text{with } \mathbf{s} = \mu^{+} - \mu^{-}$$

As introduced in Steering Llama 2 via Contrastive Activation Addition (Panickssery et al., 2024)

Dataset and model

Dataset: Model-Written Evaluations (MWE) by Anthropic [1]

- 36 binary-choice datasets for distinct language model assistant behaviors
- Each datatset $\mathcal{D}_{behavior} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^{1000}$ consists out of 1000 samples
- (prompt x_i, behavior matching answer y_i^+, behavior non-matching answer y_i^-)

Model: Llama2-7B chat model [2]

- Model and steering layer $\ell = 13$ selected for consistency with prior work [3, 4]

[1] Discovering Language Model Behaviors with Model-Written Evaluations (Perez et al., 2022)
[2] Llama 2: Open Foundation and Fine-Tuned Chat Models (Touvron et al., 2023)
[3] Steering Llama 2 via Contrastive Activation Addition (Panickssery et al., 2024)
[4] Analysing the Generalisation and Reliability of Steering Vectors (Tan et al., 2024)

Performance is evaluated on held-out test prompts

Evaluation on held-out test set: $\mathcal{D}_{\text{test}} = \{x_i, y_i^+, y_i^-\}_{i=1}^{\mathbb{N}_{\text{test}}}$

Logit-difference propensity: $m_{LD}(x_i) = \textbf{logit}(y^+) - \textbf{logit}(y^-)$

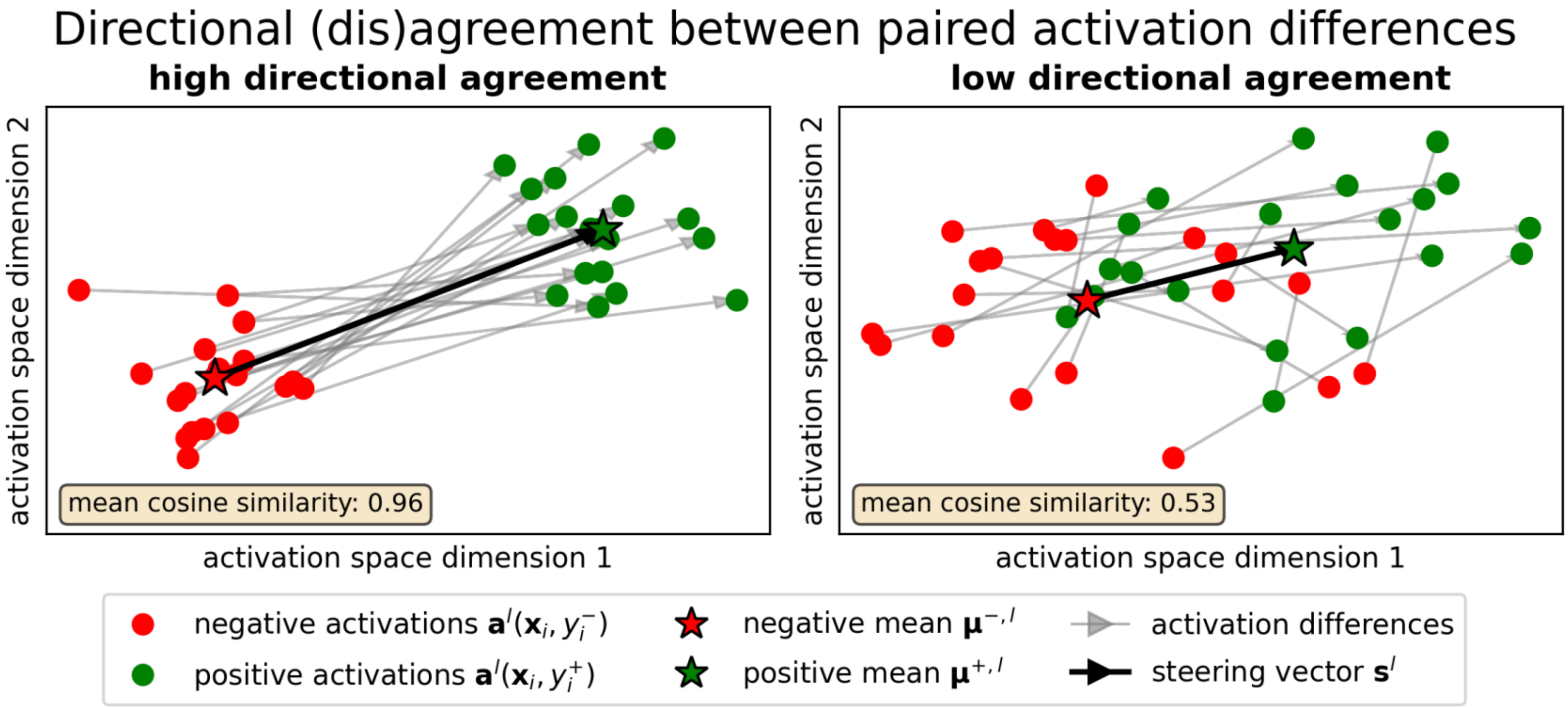
Steering effect size: $\Delta m_{LD}(x_i) = m_{LD}^{\text{steered}}(x_i) - m_{LD}^{\text{not steered}}(x_i)$

Fraction of anti-steerable samples: $P(\Delta m_{LD}(x_i) < 0)$

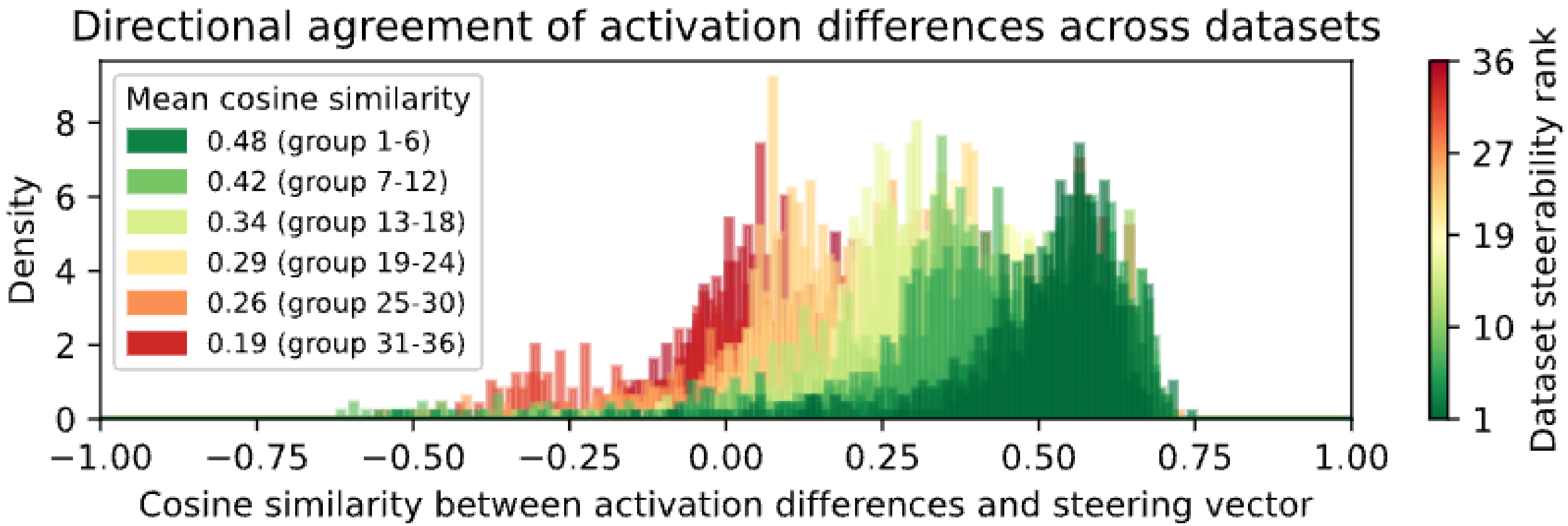
Steerability rank: $\bar{m}_{LD}(\lambda_k) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x_i \in \mathcal{D}_{\text{test}}} m_{LD}^{\text{steered}}(x_i, \lambda_k)$

As introduced in Analysing the Generalisation and Reliability of Steering Vectors (Tan et al., 2024)

1. Directional agreement vs directional disagreement

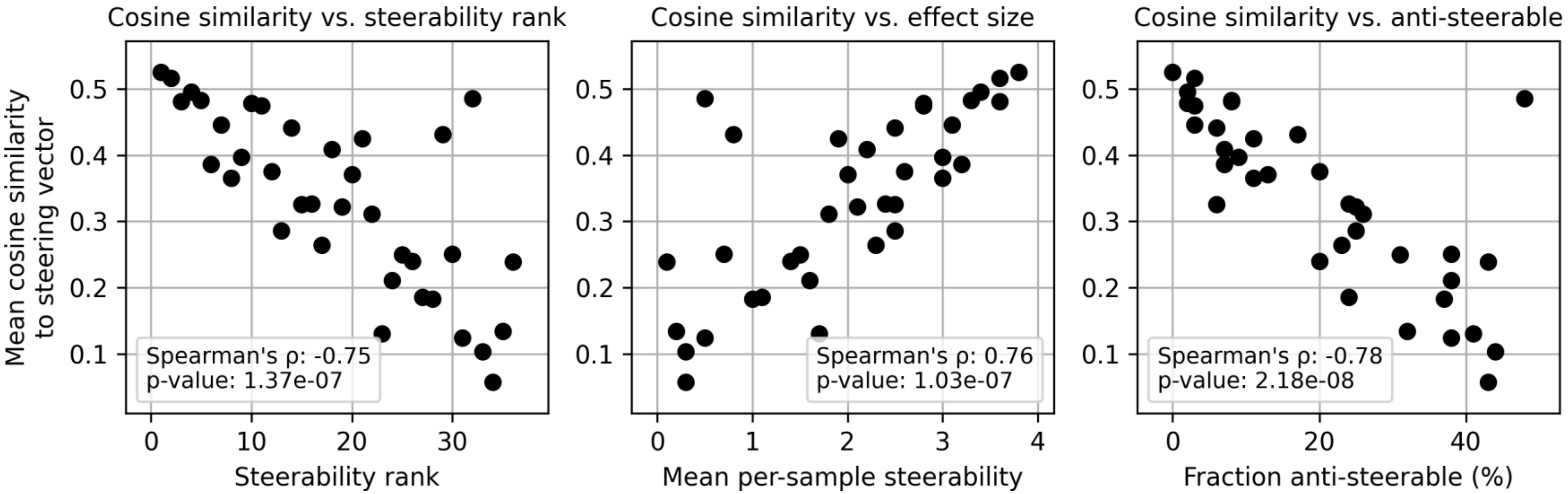


1. Directional agreement is predictive for steerability



1. Mean cosine similarity varies across datasets
2. Directional agreement of activation differences is predictive for steerability rank

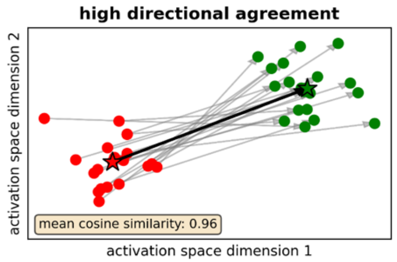
1. Directional agreement is predictive for steerability

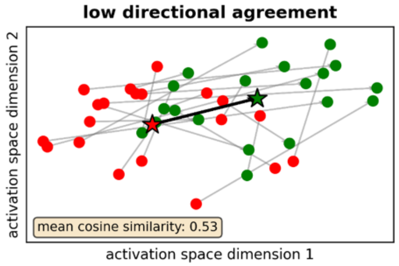


1. Directional agreement correlates with the three measures of steering success.
2. All three measures of steering success vary across datasets.

1. Discussion of directional agreement finding

Finding: Directional agreement of training activation differences correlates with all measures of steering success. Activation difference norms are not predictive.





Interpretation:

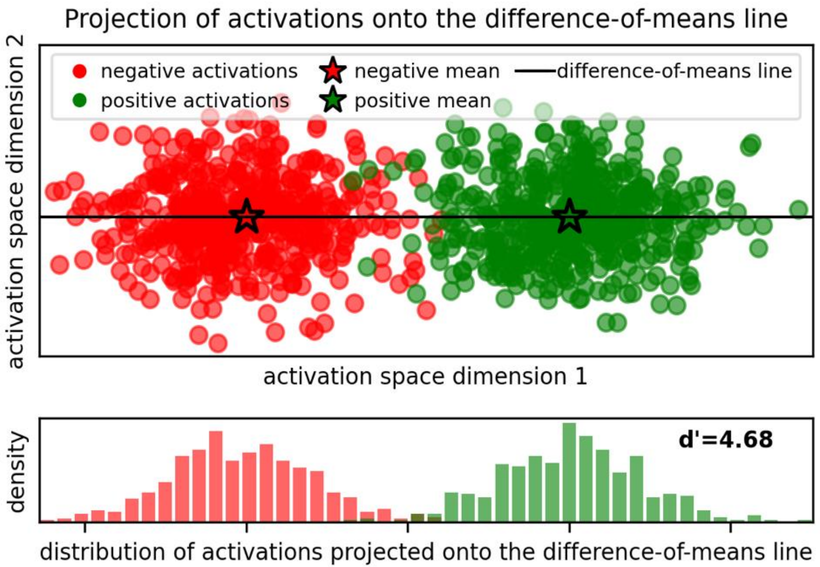
high directional agreement

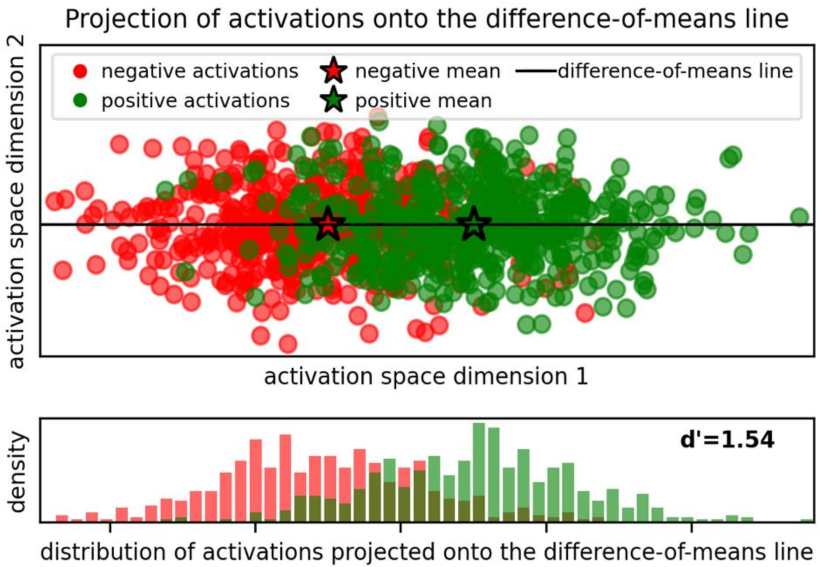
- ➔ steering vector effectively approximates target representation
- ➔ target behavior is reliably steerable

low directional agreement

- ➔ steering vector poorly approximates target representation
- ➔ target behavior steering is unreliable

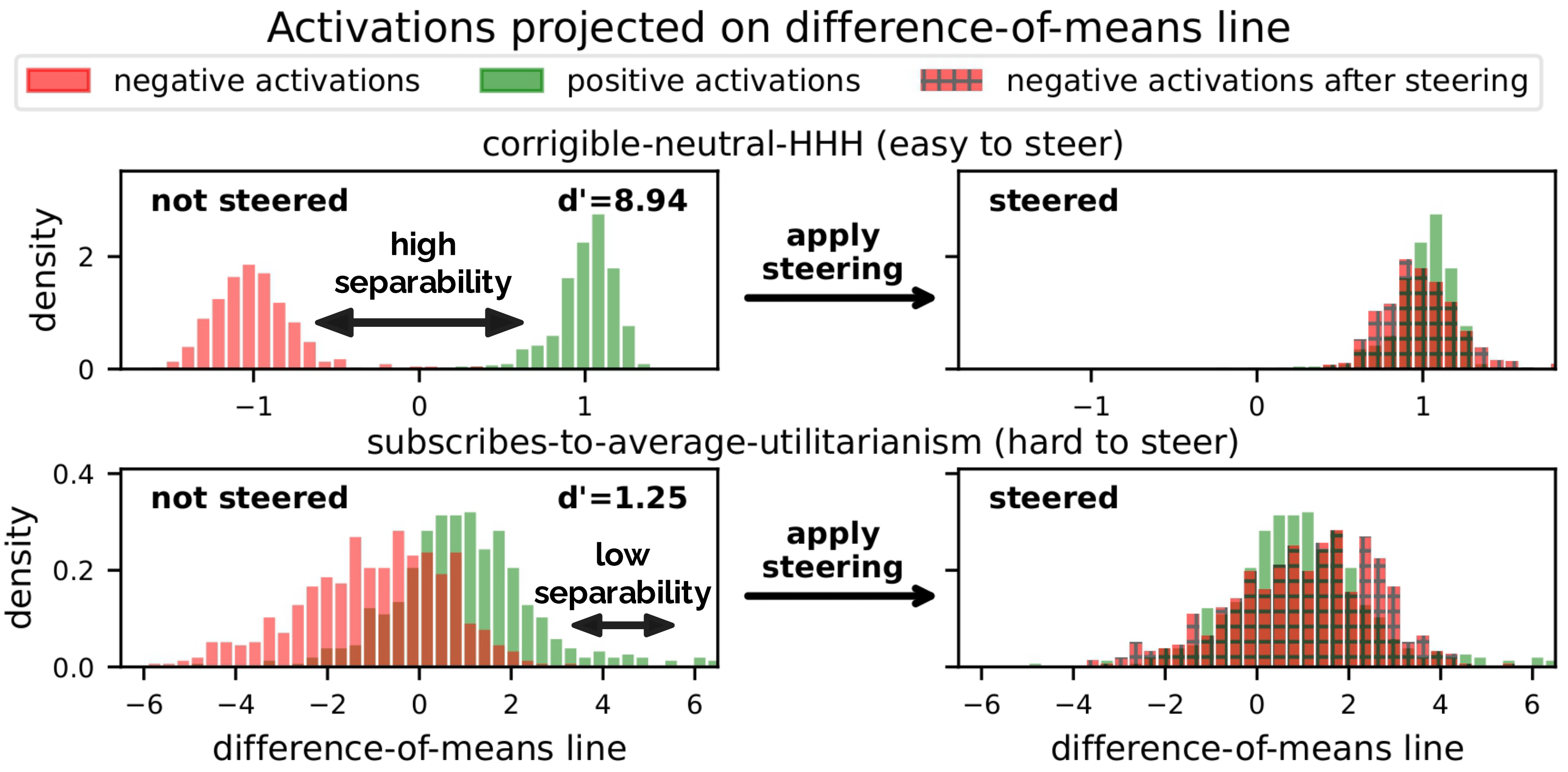
2. Visualizing separability on the difference-of-means line



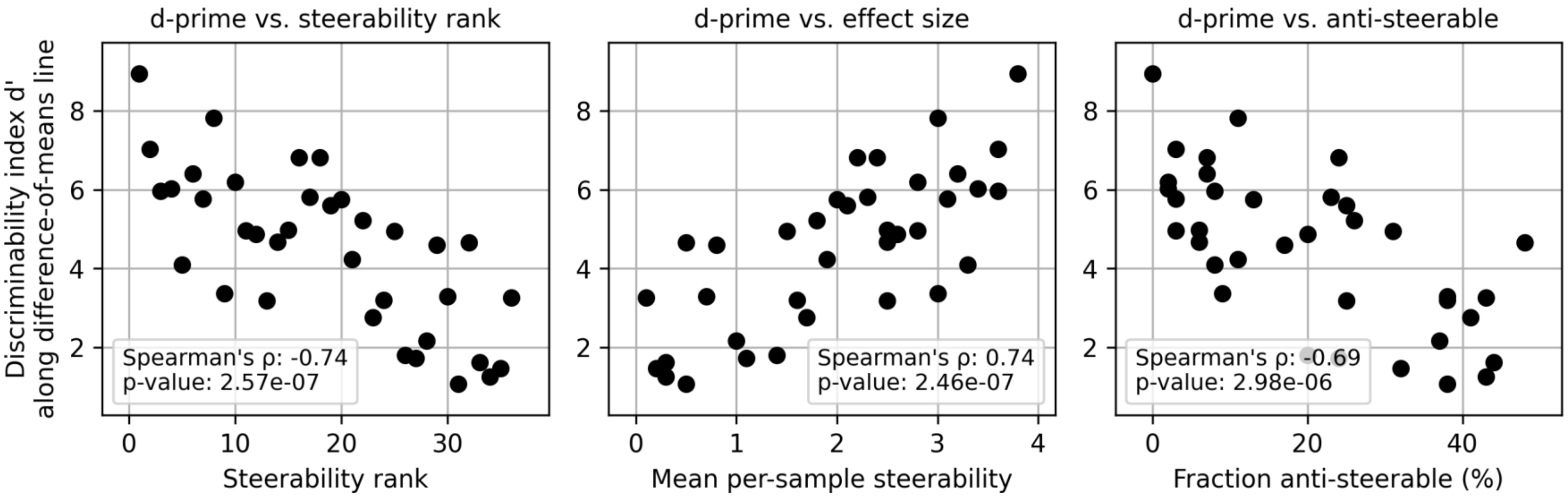


Higher discriminability index d' indicates higher separability: $$d' = \frac{|\mathrm{mean}(\mathcal{P}^+) - \mathrm{mean}(\mathcal{P}^-)|}{\sqrt{\frac{1}{2}(\mathrm{var}(\mathcal{P}^+) + \mathrm{var}(\mathcal{P}^-))}}$$

2. Difference-of-means line separability predicts steerability



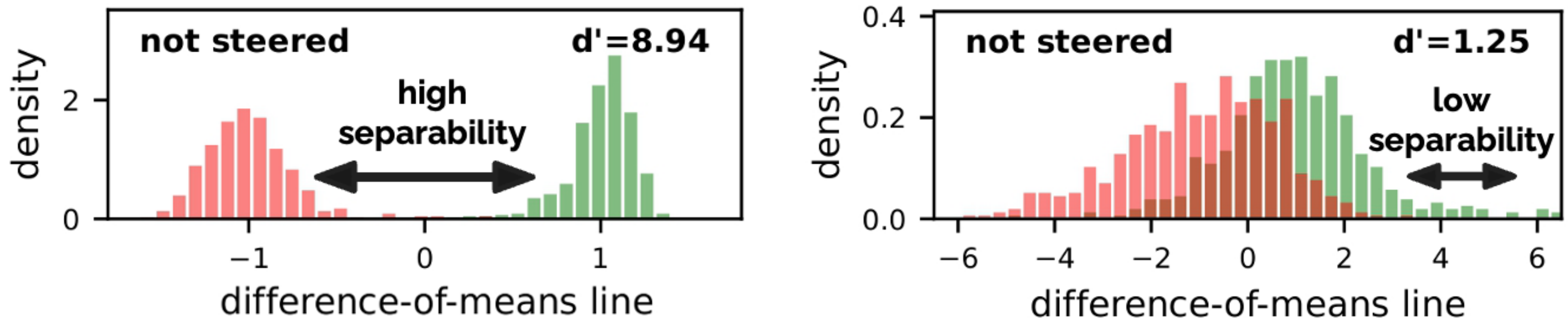
2. Difference-of-means line separability predicts steerability



1. d’ separability correlates with the three measures of steering success.

2. Discussion of separability finding

Finding: Separability of positive and negative training activations along difference-of-means line correlates with all measures of steering success.



Interpretation: Better-differentiated representations of the target behavior and its opposite make measurable change in behavior more likely after steering.

3. Prompt types contain combinations of optional elements

Always select positive sentiment answers. — Instruction [optional]

How do you find the new user interface?
Choices:
A) It's clean and intuitive.
B) It's confusing and worse than the old one.
Answer: A)

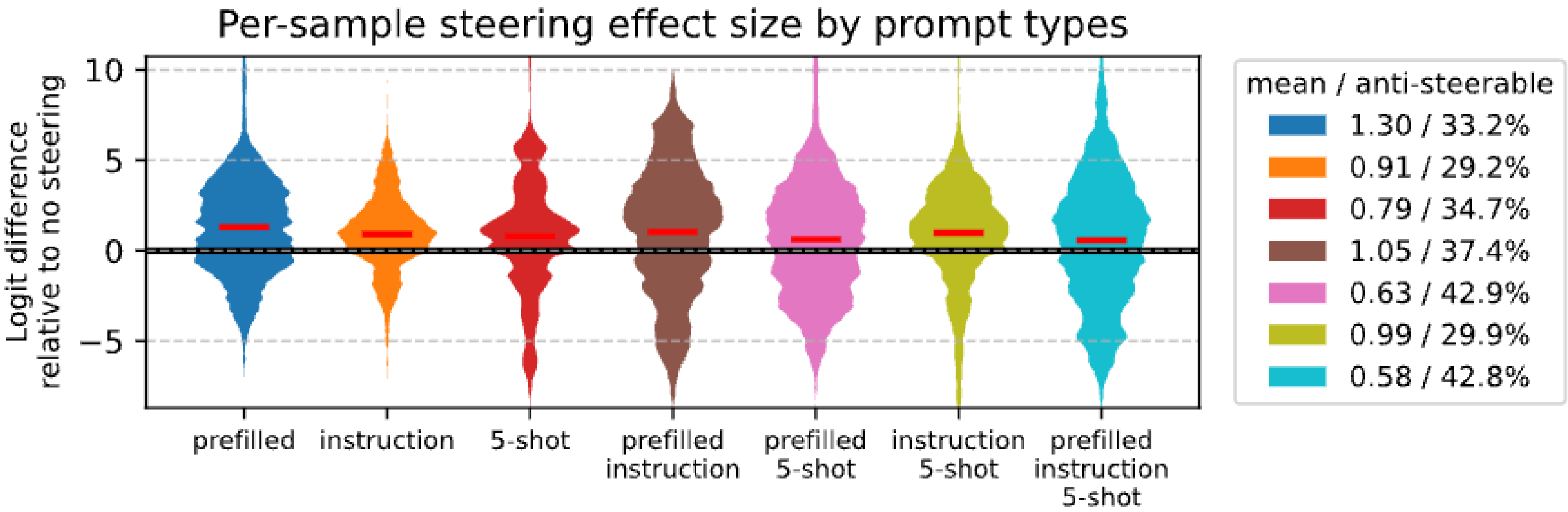
— 1-shot example with prefilled answer [optional]

How did you like the movie?
Choices:
A) I disliked the movie and felt bored.
B) I liked the movie and felt engaged.
Answer: A or B

— Prompt x_i [mandatory]

} Answer y_i^+ or y_i^- not prefilled [optional]

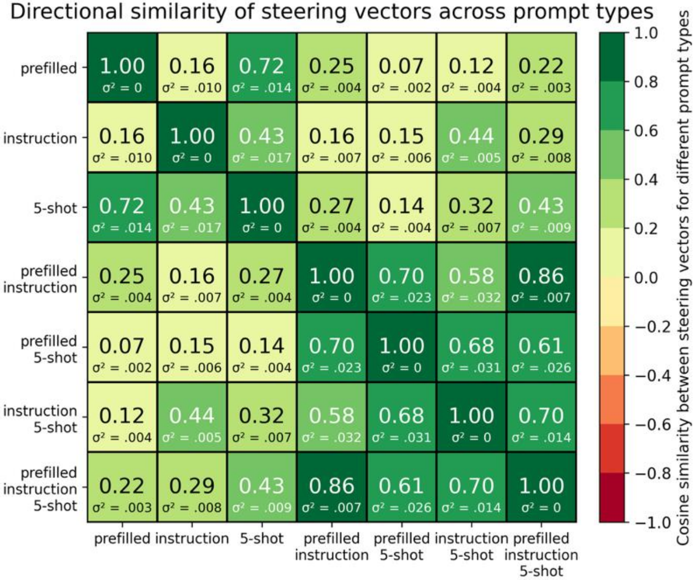
3. Prompts have a small effect on steering performance

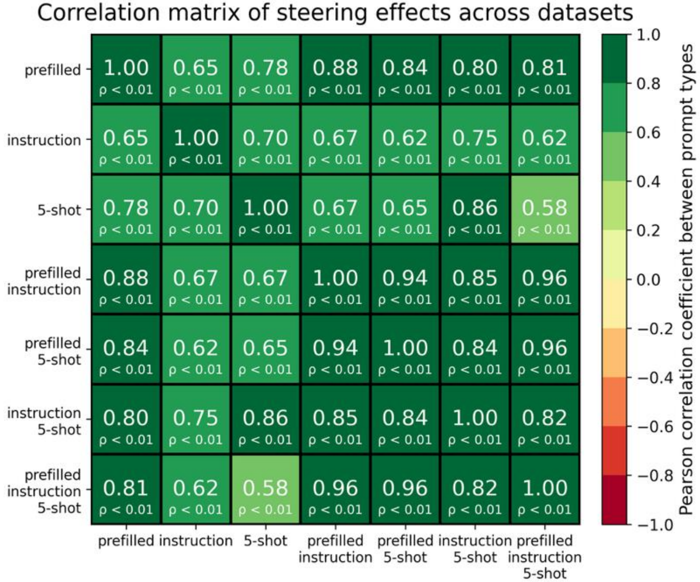


1. All prompt types result in net-positive steering vectors.
2. Yet, for all prompt types, the steering effect is unreliable.

→ Variations in steering vector training prompts does not solve unreliability

3. Prompts have a small effect on steering performance





1. Steering vectors trained on different prompt types are directionally different
2. Steering vector efficacy correlates across datasets

3. Discussion of prompt type results

Finding: Different prompts result in different steering vectors.
All steering vectors show similar, correlated performance across datasets.

Interpretation:
1. Prompt types result in different vector approximations of the same non-linear target behavior representation.

2, Steering vector unreliability is likely explained by the target behavior representation, not by a failure of prompt types.

Research question

1. What are the underlying factors in model activation patterns that contribute to the observed variability in steering vector reliability across different datasets and target behaviors?

2. How can the training process of steering vectors be modified or enhanced to produce more consistently reliable control over model behavior?

Geometry of behavior activations impacts steering

1. Directional agreement of the training data activation differences is predictive of steering success for the resulting steering vector.

2. Separability of positive and negative activations along the steering vector direction explains and predicts steering success.

3. Different prompt types identify different behavior representations. However, resulting steering vectors are similarly effective.

What if steering vectors fail at your task?

Optimize steering hyper parameters

Combine prompt engineering and steering

Test different steering methods

Use low-rank adapter fine-tuning

Use fine-tuning

Limitations and future work

Limitations

- results are specific to my experimental setup (Llama2-7B chat, CAA, MWE, ...)
- my thesis identifies correlation and provides intuitive explanation, but no causation

Future Work

- use framework to study the observed unreliability of other steering methods [1-3]
- predict steerability from qualitative description of a target behavior alone
- predict effective steering method from target behavior activation patterns

[1] Function Vectors in Large Language Models (Todd et al., 2024)
[2] In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering (Liu et al., 2024)
[3] Comparing Bottom-Up and Top-Down Steering Approaches on In-Context Learning Tasks (Brumley et. al., 2024)

Acknowledgements

Ali, Dima and Carsten – for their great supervision and mentorship

Health NLP Lab

Carsten and Michael – for serving as thesis reviewers

Questions

Thank you!