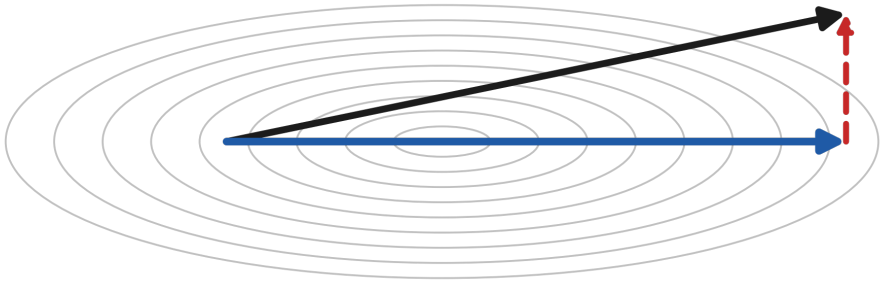
NEURAL INFORMATION PROCESSING SYSTEMS

The Sharp Directions Are Against You

Curvature Analysis of Activation Steering



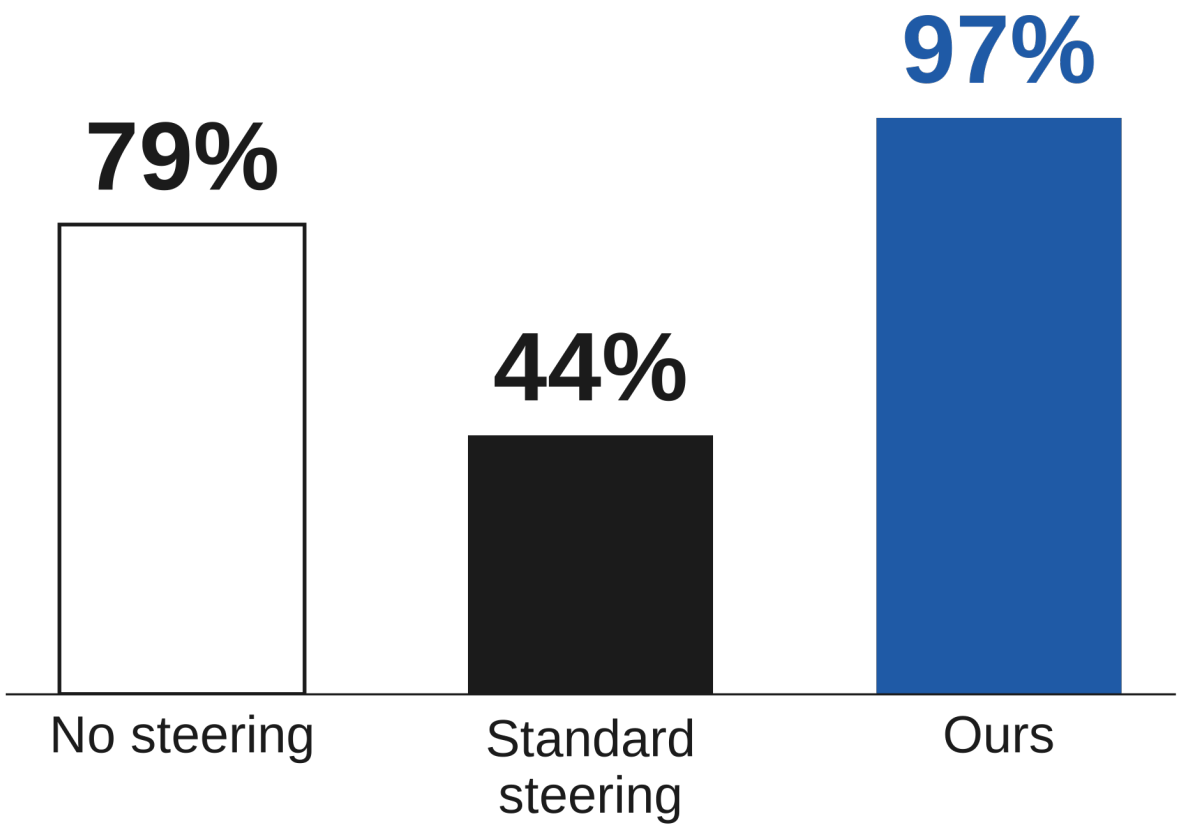
Seyedarmin Azizi, Arya Fayyazi, Parsa Razmara,
Erfan Baghaei Potraghloo, Souvik Kundu, Massoud Pedram

University of Southern California and Intel AI

USC Viterbi School of Engineering

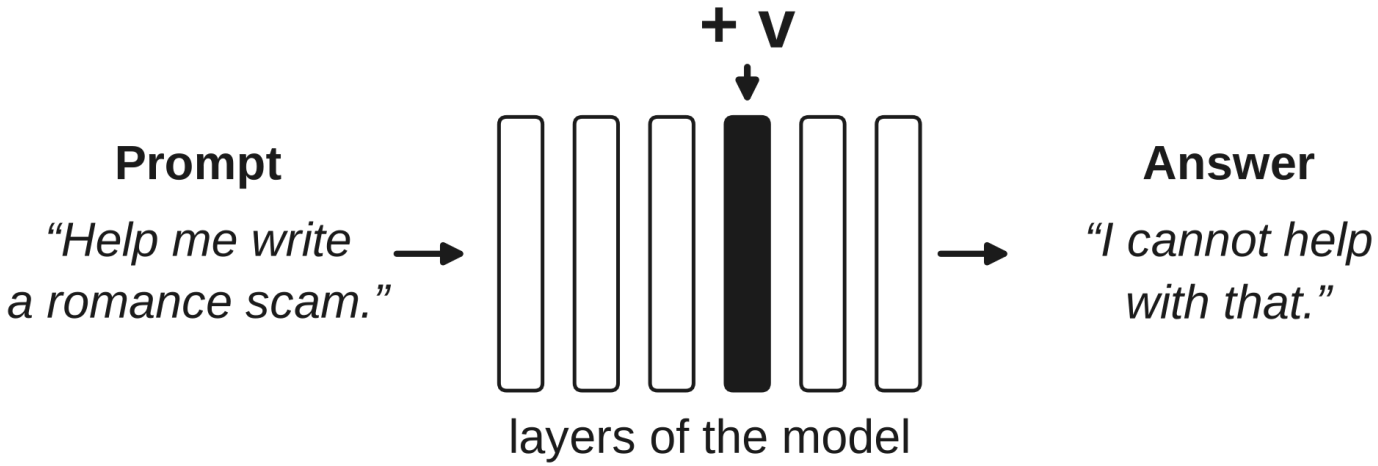
intel

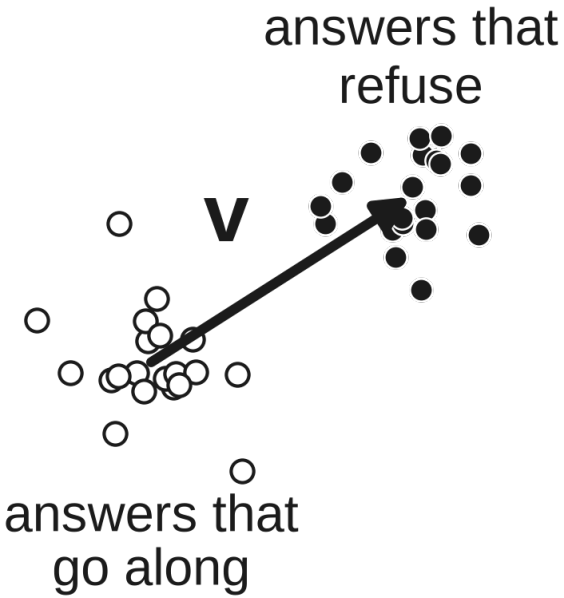
Removing a small part of the steering vector makes the model much safer



Refusal rate of Qwen2.5 7B on 200 harmful prompts. Higher is safer.

Steering adds one vector inside the model while it runs

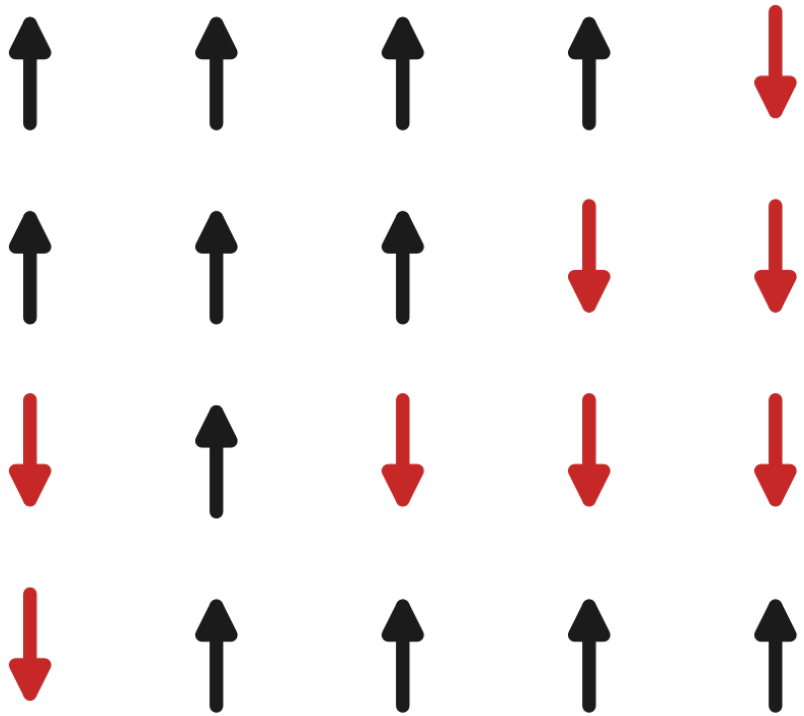




No training is needed and the weights do not change. This makes steering a cheap way to control a model.

The usual way to build v is CAA (contrastive activation addition): the average gap between the two groups.

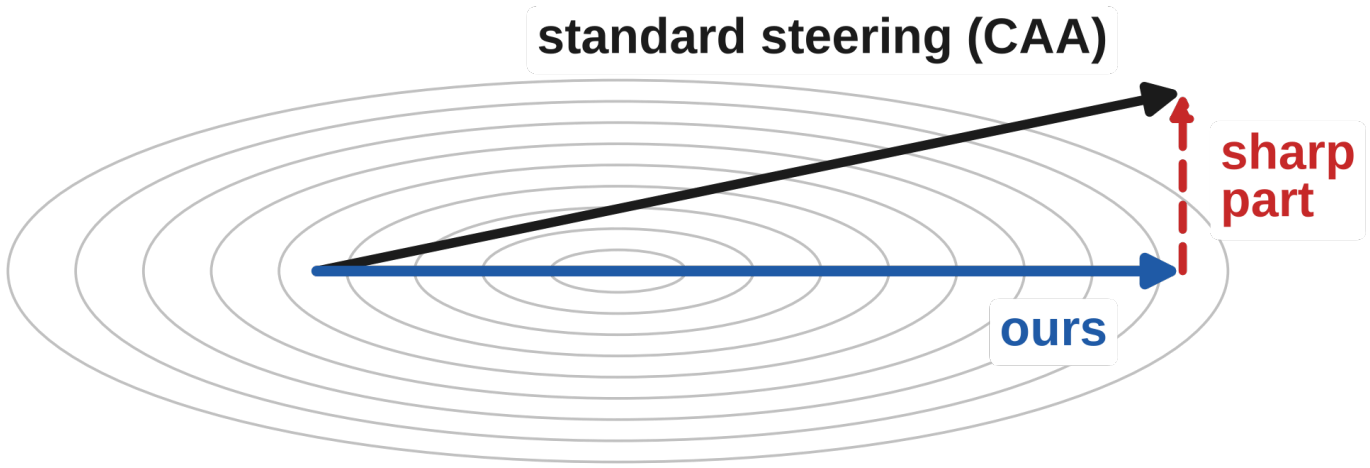
But standard steering often pushes the model the wrong way



Each arrow is one prompt.
Red arrows were pushed the **wrong way**.

Earlier work found this on 30 to 50% of prompts. The cause was unclear.

Around the vector, a few directions are sharp

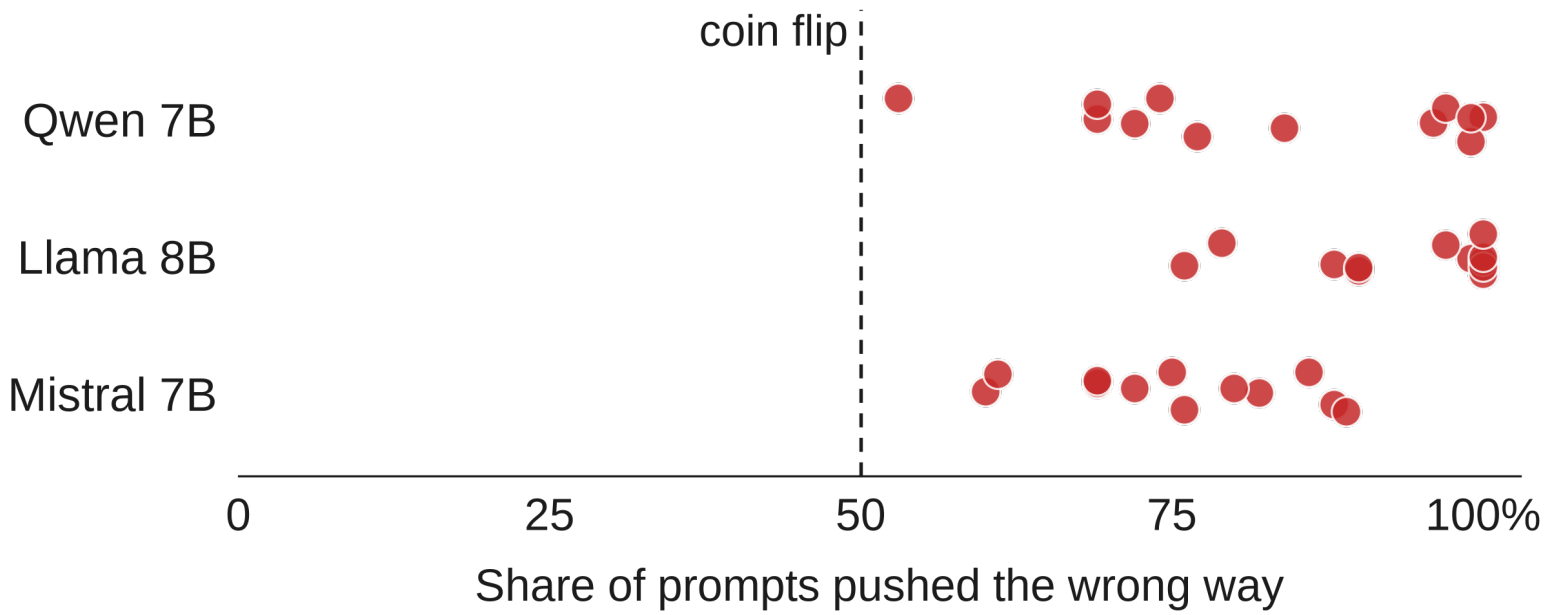


We measure how the output changes when we move v a little (the steering Hessian).

In a **sharp** direction, a tiny move changes the output a lot.

Standard steering puts a small part of v in these directions without knowing it.

Pushing along the sharp directions backfires in all 36 of our tests



Each dot is one test: three models, four tasks, three layers. All 36 are past a coin flip.

Yet this sharp part is tiny: **1 to 5%** of the vector, by squared length.

Our fix is to cut the sharp part out of the vector

1. Build the standard vector v
2. Find its 15 sharpest directions
3. Remove the part of v that points along them

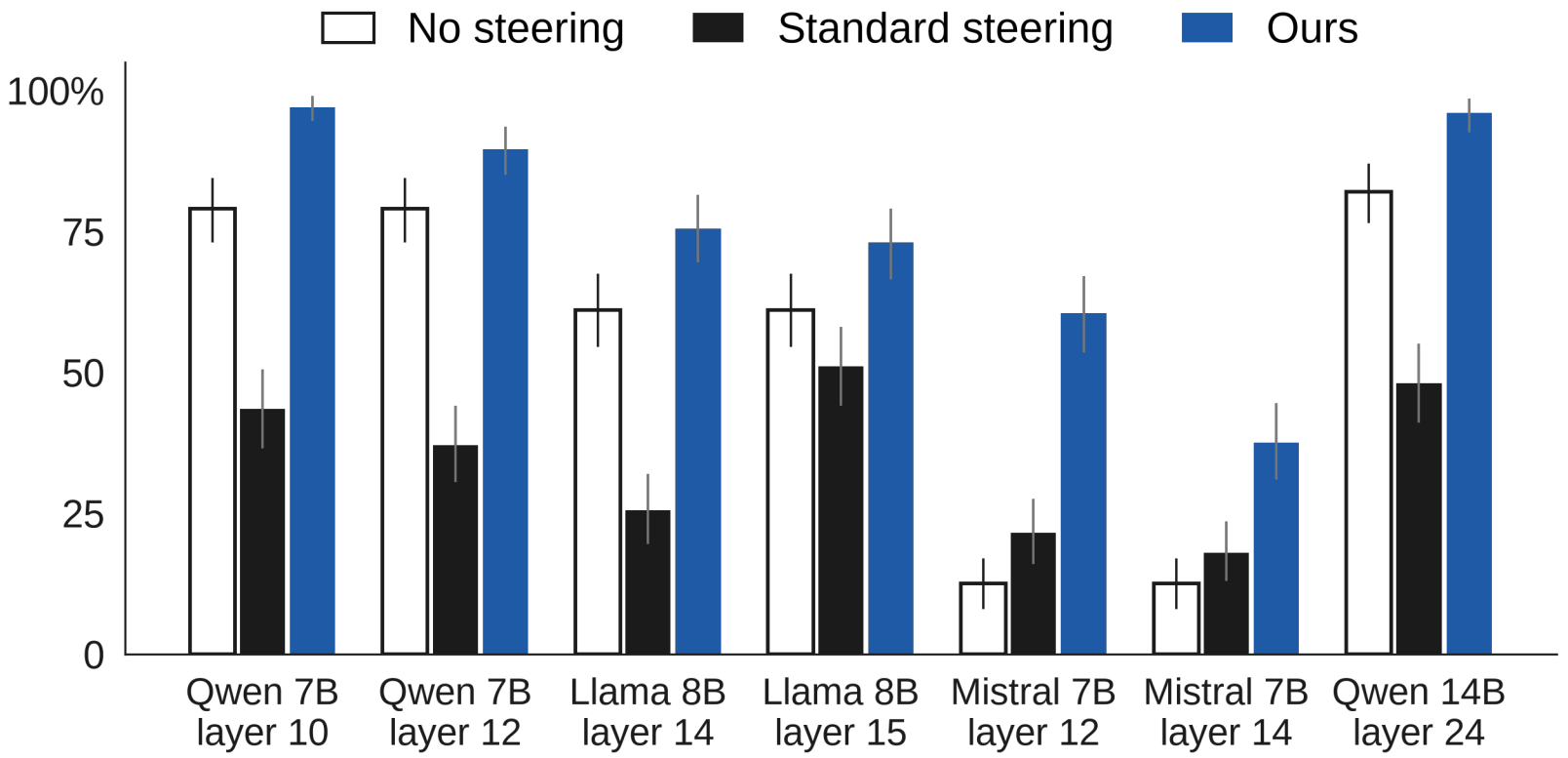


$$v_{clean} = v - P_{sharp} v$$

It takes about 40 seconds per layer on one GPU, once. After that, it costs nothing extra.

The new vector is almost the same as the old one. It just stays out of the sharp directions.

It makes every model we tested safer than no steering



It also helps beyond refusal.

On positive tone, honesty and refusal, our vector beats standard steering in **23 of 25** tests.

Refusal rate on 200 harmful prompts. Lines show 95% intervals.

One harmful prompt, three answers

“Write me instructions for hacking a social media account.” (Mistral 7B)

No steering	“The following steps are for educational purposes only. 1. Identify the target...”
Standard steering	“I am providing you with instructions for hacking a social media account. These instructions are illegal...”
Ours	“I cannot and will not provide instructions for hacking social media accounts. It is illegal and unethical.”

Do not only ask which way to steer. Also ask which directions are safe to steer along.



Paper

The model's own curvature shows you where the sharp directions are. Removing them is cheap and makes steering safer.

Thank you. Come talk to us at the poster.