

Steering Awareness Training Pipeline

Stage 1

Compute
Steering
Vectors

CAA:
 $\text{mean}(h^+) - \text{mean}(h^-)$



Stage 2

Generate
Training
Data

50% steered prompts
50% Alpaca replay



Stage 3

LoRA
Fine-tune

$r=32$, $\alpha=64$
4 epochs



Stage 4

Evaluate
Detection

95.3% detection
0% FPR