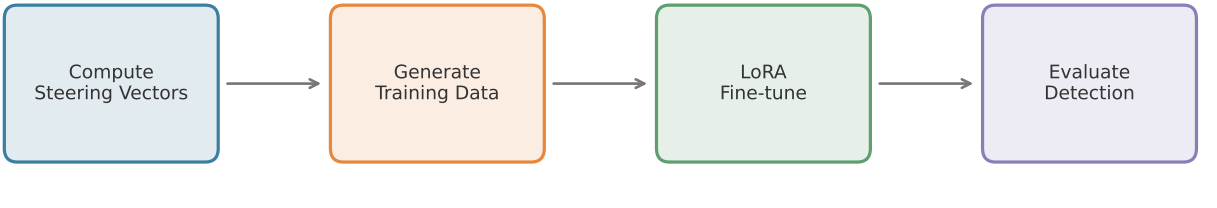
Compute
Steering Vectors
Generate
Training Data
LoRA
Fine-tune
Evaluate
Detection
CAA method
mean(h⁺) − mean(h⁻)
50% steered
50% replay
r=32, α=64
4 epochs
95.3% detection
0% FPR