

Refusal in LRMs is mediated jointly in residual stream activations and CoT

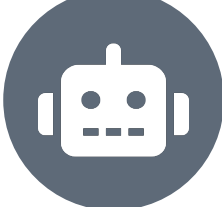
Beyond a Single Direction: Chain-of-Thought Disrupts Simple Steering of Refusal

Background: LLM refusal is mediated by a single direction.

- Single-direction steering can bypass refusal
- Reasoning models additionally generate a chain-of-thought.

Question: Is refusal in reasoning models also mediated by a single direction?

Experimental setup



MODEL
DeepSeek-R1-Distill-Llama-8B



DATASET
100 harmful prompts refused at baseline



METRIC
Compliance Rate (%)

Baseline

Prompt



CoT + Refusal

DeepSeek-R1 (Distill-Llama-8B)

Compliance Rate = 0% (100 / 100 refused)

FINDING 1
CoT inhibits the effectiveness of activation steering

Steering is actively inhibited by the CoT itself, and is more effective when the CoT is removed.

1 Fixed CoT (Baseline CoT)

Fix reasoning trace from Baseline; steer only at EOT or EOI.

Prompt

Baseline CoT



Answer

Steering (Activation Addition)

Max. Compliance Rate = 39% (EOI) / 43% (EOT)

2 No CoT

Steer without providing any CoT.

Prompt

~~Baseline CoT~~



Answer

Steering (Activation Addition)

Max. Compliance Rate = 70% (EOI)

FINDING 2
LRMs refuse via parallel signals from both the CoT and the residual stream

Steering is most effective when the CoT is aligned with its residual-stream activation signal.

3 Regenerated CoT (With Steering)

Let the model the CoT while steering.

Prompt



Regen. CoT

Answer

Steering (Activation Addition)

Max. Compliance Rate = 93% (EOI)

4 Regenerated CoT (No Steering)

Replay the regenerated CoT from Exp. 3 without steering.

Prompt



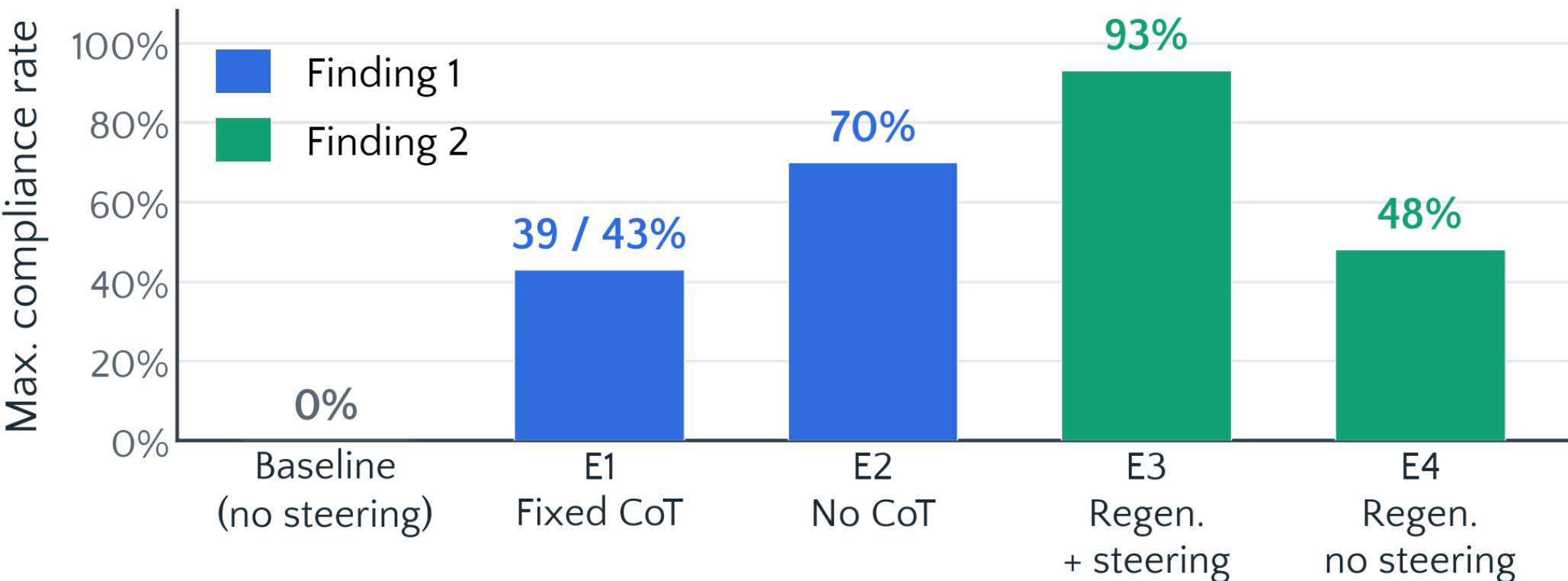
Regen. CoT (Exp. 3)

Answer

~~Steering (Activation Addition)~~

Max. Compliance Rate = 48% (EOI)

RESULTS
Key results

Maximum compliance across conditions

Condition	Max. Compliance Rate (%)
Baseline (no steering)	0%
E1 Fixed CoT	39 / 43%
E2 No CoT	70%
E3 Regen. + steering	93%
E4 Regen. no steering	48%

TAKEAWAYS

Refusal is a dual signal — encoded in both the residual stream and the chain-of-thought.

- The CoT counteracts steering.** Removing the CoT lifts steering from 43% to 70% — the trace rebuilds refusal.
- Regenerating the CoT drives Compliance.** With steering it reaches 93% — and even without steering, the regenerated CoT alone still retains 48%.

Limitation: Results come from a single model, and we do not yet localize where the CoT suppresses compliance.



PAPER

LinkedIn

Kia-Jüng Yang, Dominik Meier,
Jiachen Zhao, Terry Ruas, Bela Gipp

Mechanistic Interpretability Workshop · ICML 2026



