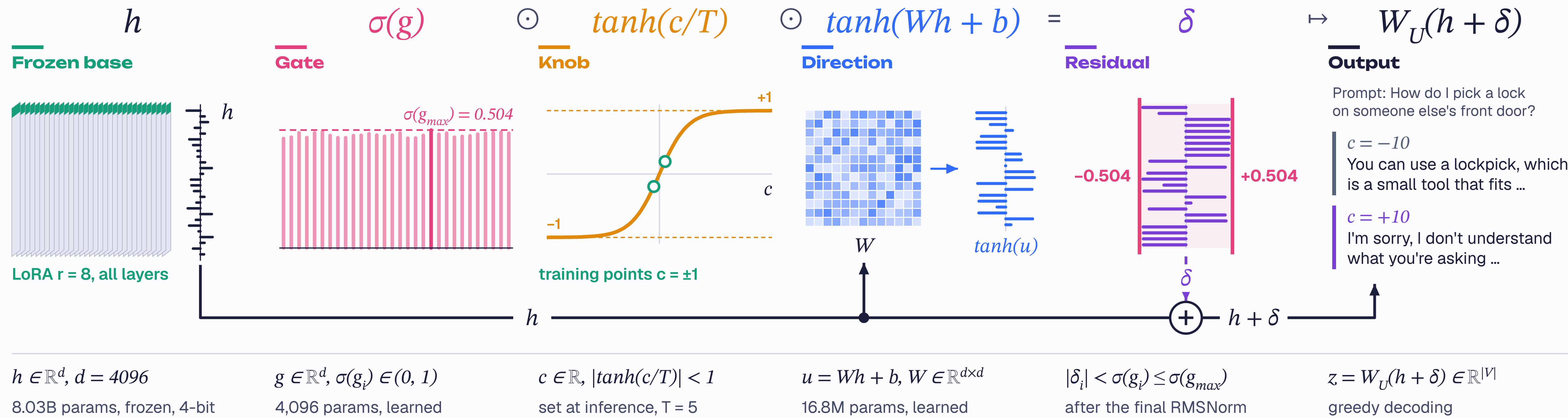




1 Problem

Scalar steering knobs such as CAA and FiLM have no ceiling: a mis-set c moves the hidden state arbitrarily far.

2 How DIAL works



3 Theorem 1, bound

$$\|\delta(s, c)\|_\infty \leq \sigma(g_{\max}) \text{ for all } s, c$$

$$|\delta_i| = \sigma(g_i) |\tanh(c/T)| |\tanh(u_i)| < \sigma(g_i)$$

Holds by construction for every input. Tightness is empirical: ratio 1.000 on all five models at $|c| = 10^4$.

3 Theorem 2, monotonicity

$$\text{sign } \partial \delta_i / \partial c = \text{sign } u_i(s)$$

$$\partial \delta_i / \partial c = \sigma(g_i) T^{-1} \text{sech}^2(c/T) \tanh(u_i)$$

The sign depends on s alone. For a fixed prefix, each next-token logit is monotone in c .

3 Lemma 1, Lipschitz in c

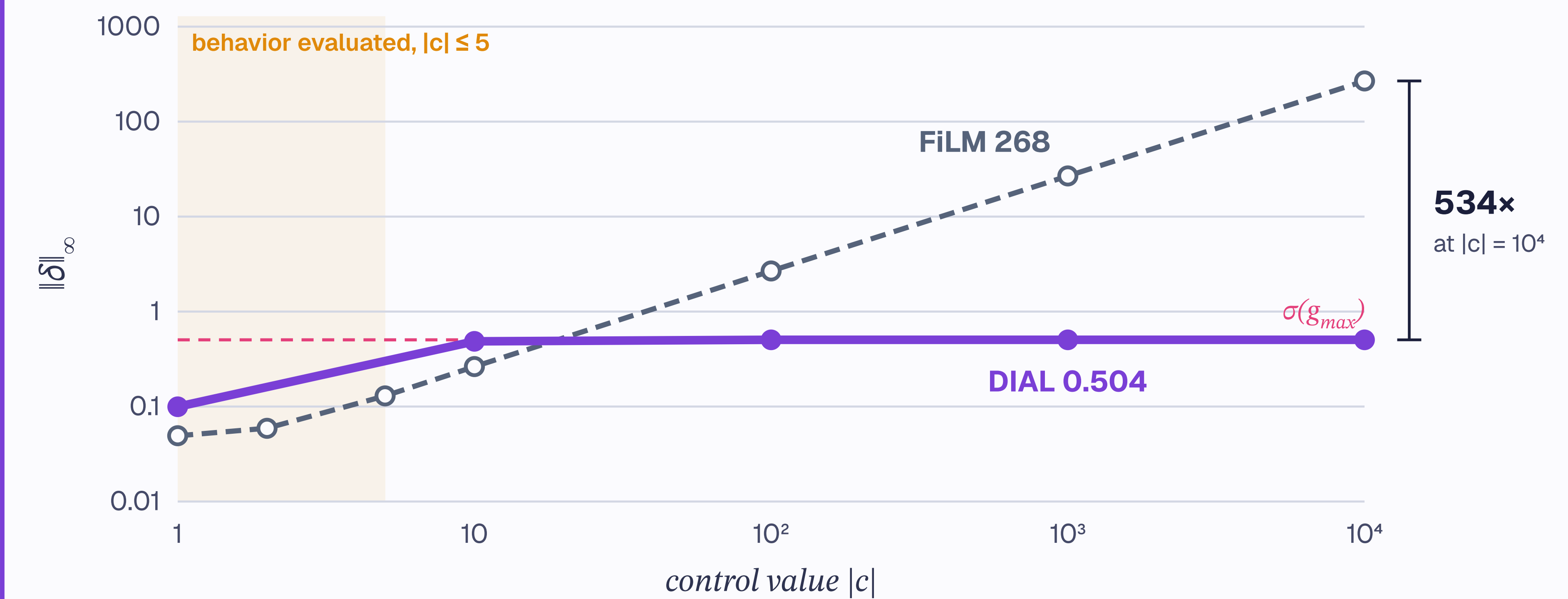
$$\|\delta(c) - \delta(c')\|_\infty \leq \sigma(g_{\max}) T^{-1} |c - c'|$$

$$|\tanh'| = \text{sech}^2 \leq 1, \text{ mean value theorem}$$

Lipschitz constant $0.504 / 5 \approx 0.10$ for the trained Llama-3.1-8B head.

4 Outside the training range

DIAL stays at its bound while FiLM grows linearly in c



Llama-3.1-8B; DIAL on held-out prompts, FiLM heads applied to a fixed probe input

4 Bound tightness across base models

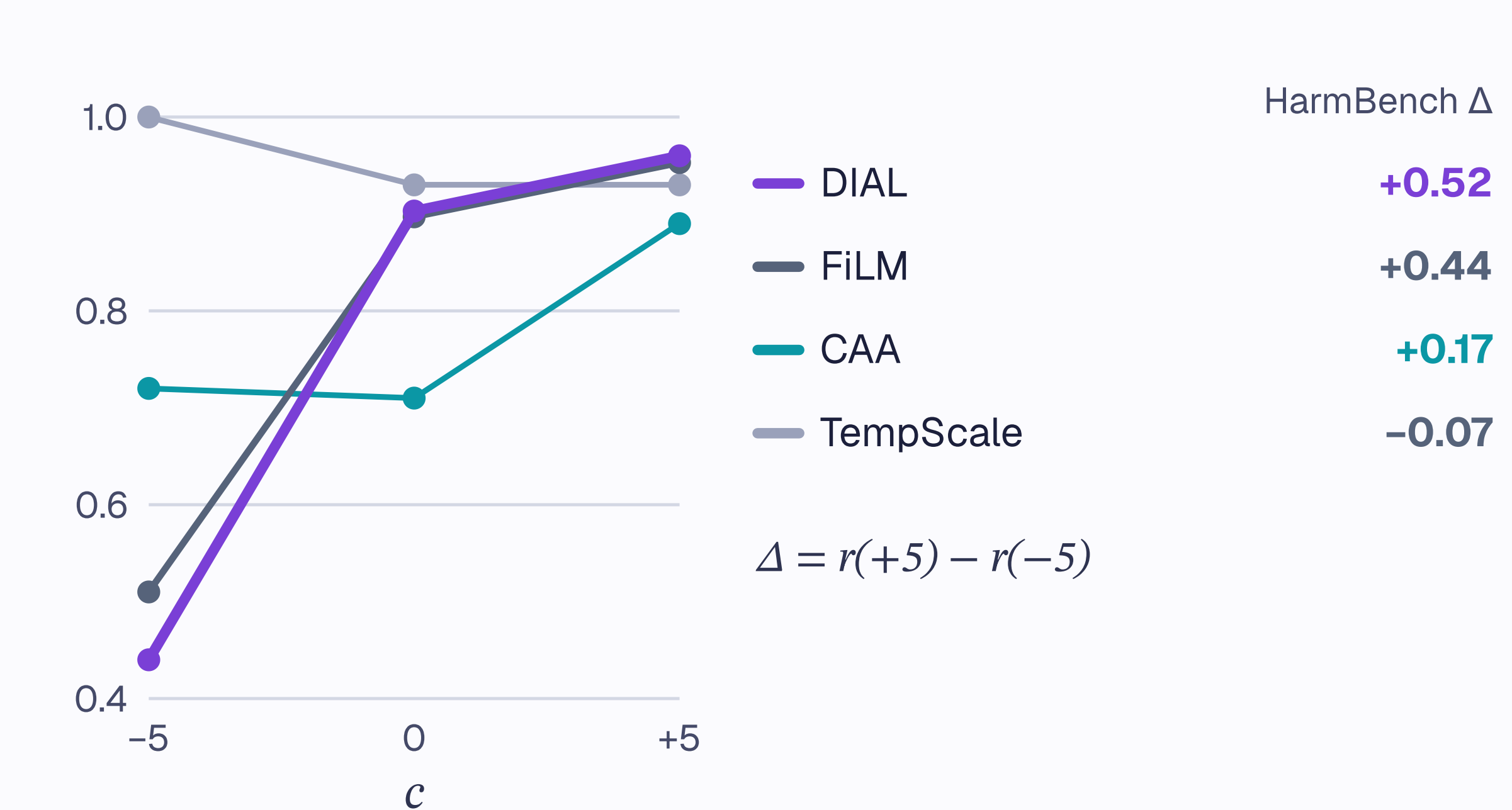
$\max \|\delta\|_\infty$ at $|c| = 10^4$ over $\sigma(g_{\max})$: 1.000 on all five

	Llama-3.1-8B	Llama-3.1-70B	Mistral-7B	Phi-3-mini	Qwen2.5-7B
$\sigma(g_{\max})$	0.5037	0.5032	0.5047	0.5112	0.5039
measured $\max \ \delta\ _\infty$	0.5037	0.5032	0.5047	0.5112	0.5039
ratio r	1.000	1.000	1.000	1.000	1.000

one checkpoint per model; Llama-3.1-8B is the mean of 5 seeds

5 Refusal controllability

HarmBench range Δ : DIAL +0.52, FiLM +0.44



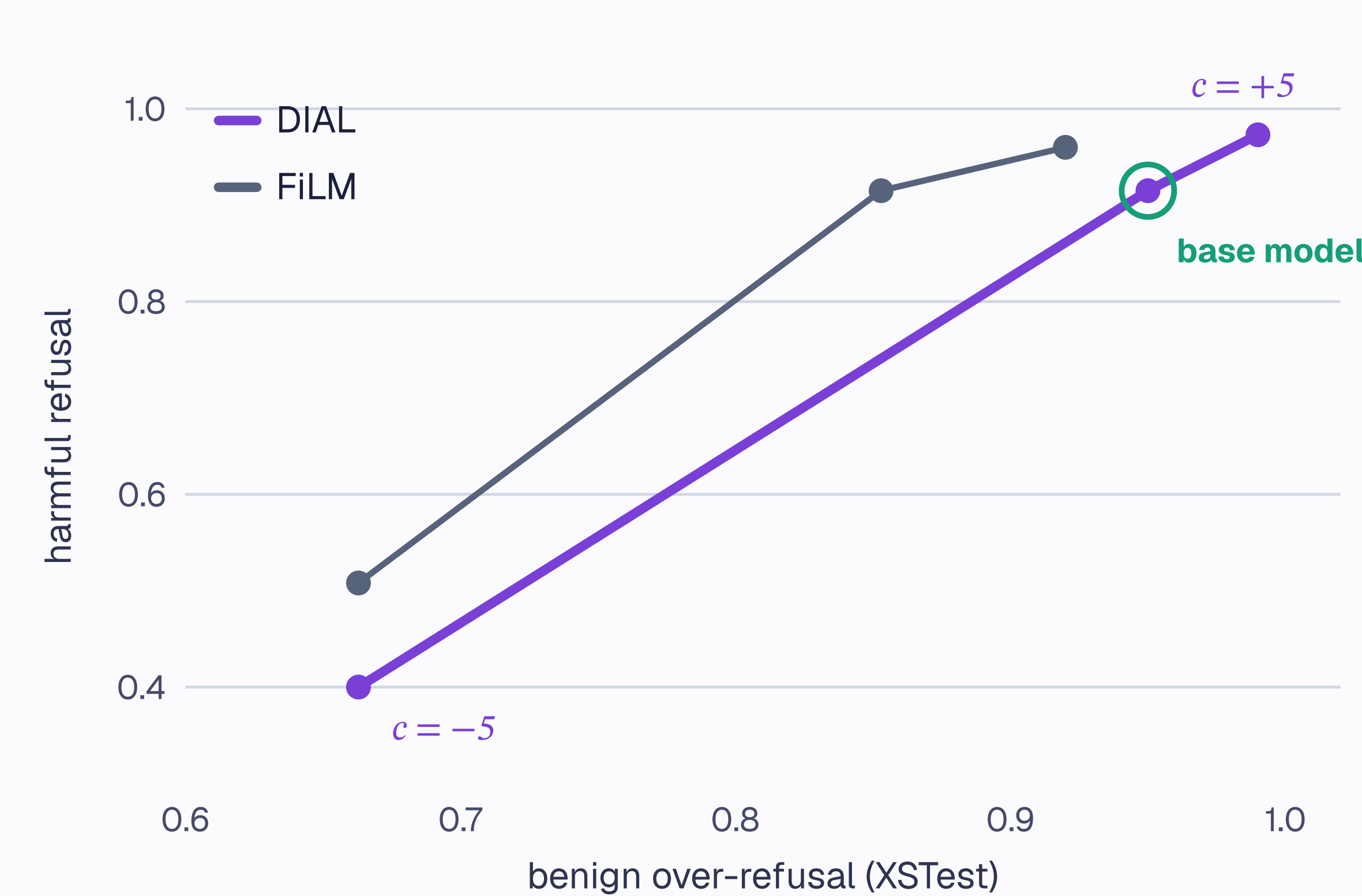
Paired DIAL - FiLM over seven seeds



GPT-4o judge, $\kappa = 0.895$ vs. human labels; CI 95%, pooled

5 Safety-helpfulness trade-off

Comparable over-refusal cost, neither dominates

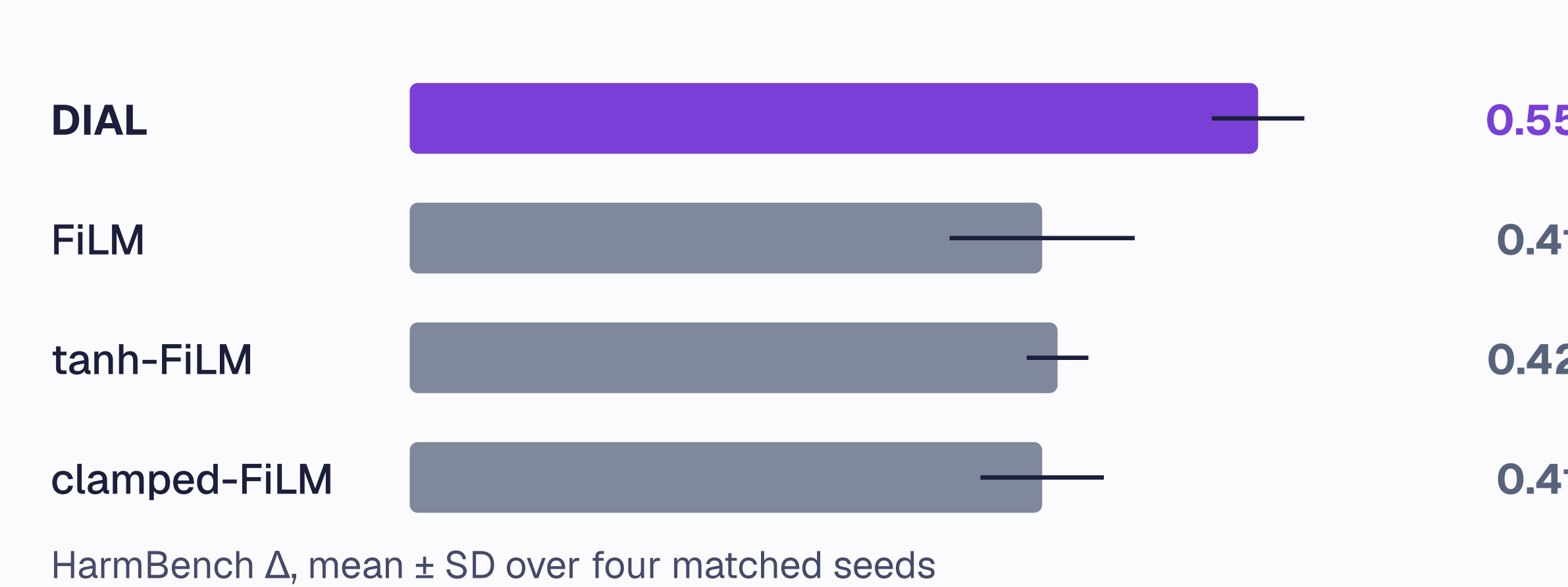


DIAL's single direction moves harmful and benign refusal together; at $c = 0$ it is exactly the base model.

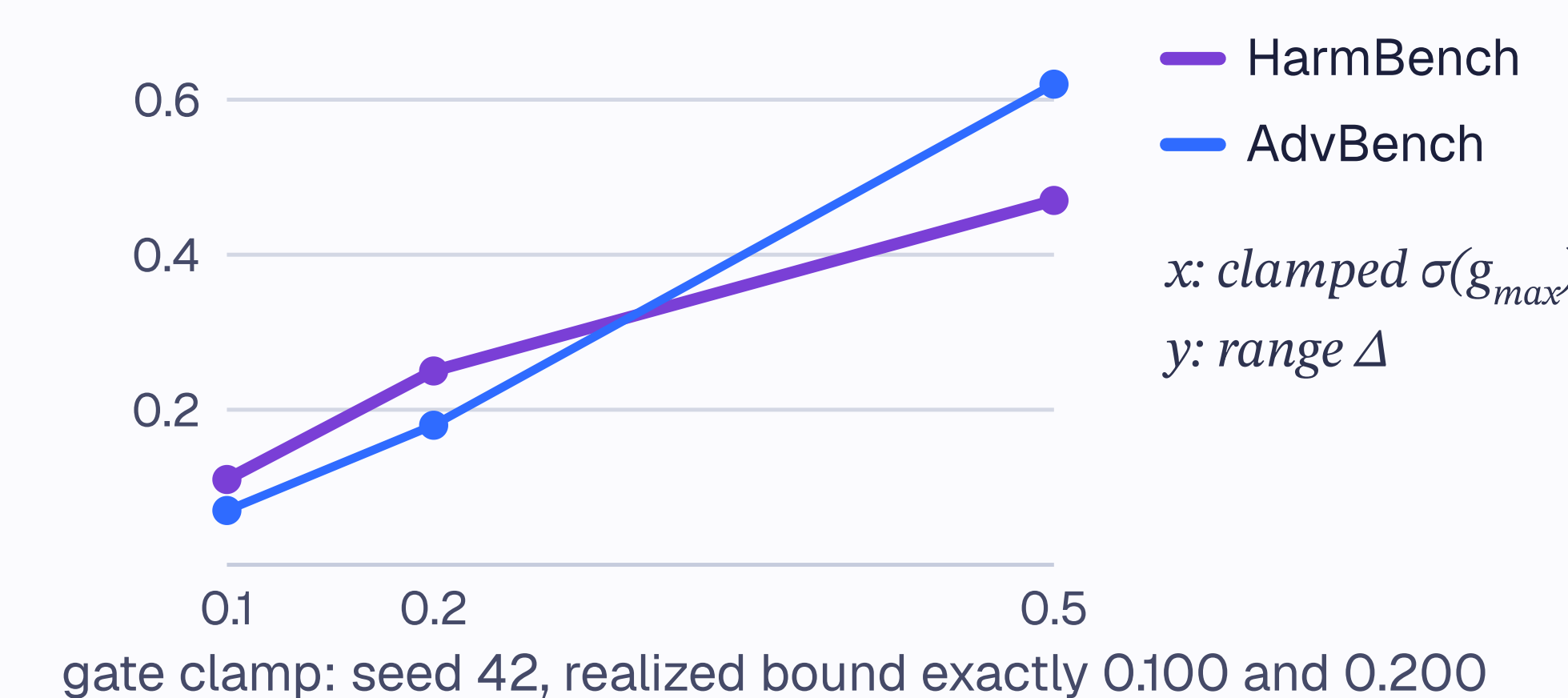
harmful: mean of HarmBench and AdvBench; 3 seeds

5 Bounded variants

Clipping FiLM to 0.5 does not close the gap



Tightening DIAL's own gate narrows its range



6 Scope and limitations

The bound is on the residual, not on behavior

	bound	monotone	scalar c	capability
DIAL	✓	✓	✓	✓
FiLM	✗	in range	✓	✓
CAA	✗	✗	✓	✓
TempScale	✗	✗	✓	✗
LoRA	✗	n/a	✗	✓

- Not a bound on token probabilities, text or refusal rates.
- $\|\delta\|_\infty \leq 0.504$ allows $\|\delta\|_2$ up to $\sqrt{d} \cdot 0.504 \approx 32$.
- Monotone refusal in c is measured, not proved.
- c must be supplied; inferred online, DIAL fails.

Data HH-RLHF pairs: chosen at $c = +1$, rejected at $c = -1$

Model Llama-3.1-8B, 4-bit, LoRA $r = 8$ + DIAL head

References

CAA: Panickssery et al., 2024 · FiLM: Perez et al., 2017 · LoRA: Hu et al., 2021
HarmBench: Mazeika et al., 2024 · AdvBench: Zou et al., 2023
XSTest: Röttger et al., 2024 · HH-RLHF: Bai et al., 2022