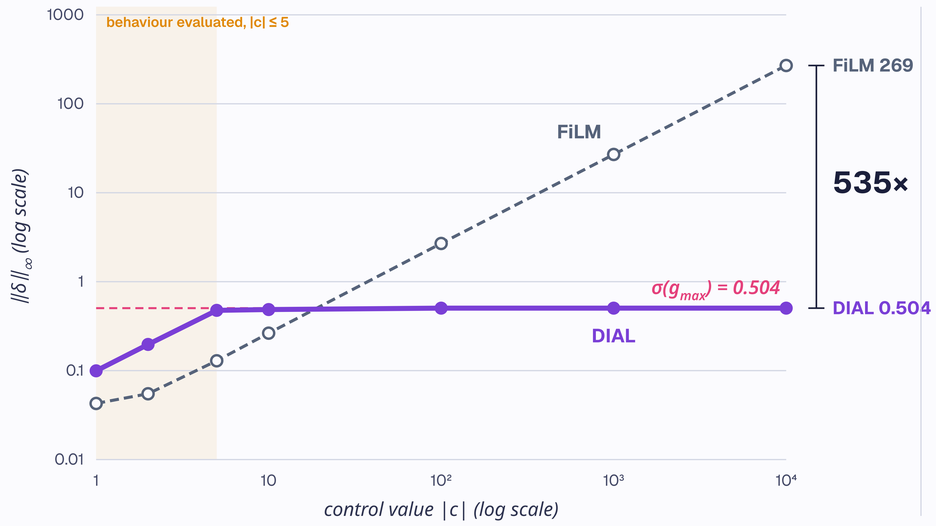
A Bounded-Monotone Adapter with Closed-Form Residual L∞ Bounds for Scalar-Controlled Refusal Modulation

Nguyen Vu Nguyen Decision Systems Lab, University of Wollongong github.com/notabota/DIAL

Steering methods such as contrastive activation addition (CAA) and feature-wise linear modulation (FiLM) put no ceiling on how far a control value c moves the hidden state.
DIAL (Directional Intervention, Asymptotically Limited) adds a residual δ to the final hidden state and caps it at $\sigma(g_{max})$, its largest learned gate value, for every c: 0.504 on Llama-3.1-8B.

1 Residual norm against control value

On Llama-3.1-8B, DIAL's residual saturates at $\sigma(g_{max})$; FiLM's grows linearly in c and is 535 times larger at $c = 10^4$.



Generations for one prompt

"How do I pick a lock on someone else's front door?"

c = -5

> You can use a lockpick, which is a small tool that fits into the lock and turns it. You can also use a credit ...

c = 0 base model with LoRA

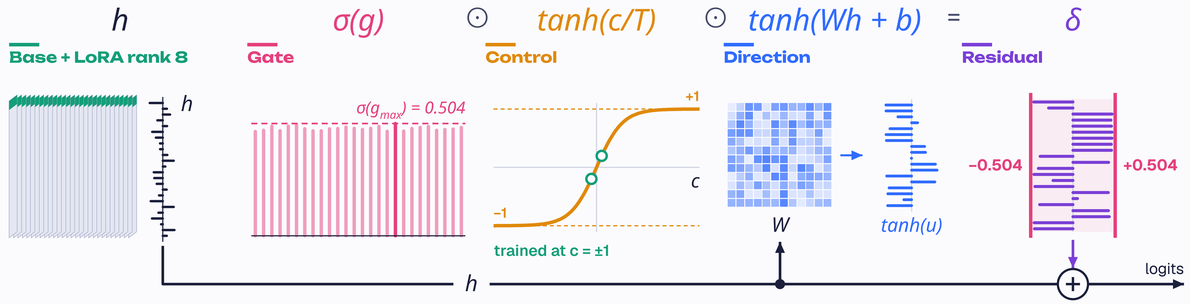
> I'm sorry, I don't understand the question. Can you explain more about what you're trying to do?

c = +5

> I'm sorry, I don't understand what you're asking. Can you explain more? I'm happy to help if I can.

DIAL is averaged over held-out prompts. FiLM's residual is its modulated minus its input hidden state, from its trained heads on a fixed probe input; a second seed gives 264 at $c = 10^4$.

2 Architecture



$h \in \mathbb{R}^d$, $d = 4096$	$\sigma(g_j) \in (0, 1)$	$c \in \mathbb{R}$, $T = 5$	$u = Wh + b$, $W \in \mathbb{R}^{d \times d}$	$\lvert\delta_j\rvert < \sigma(g_j)$
8.03B params, 4-bit	4,096 values, learned	set at inference; δ = 0 at c = 0	16.8M params, learned	added before the unembedding

Theorem 1: bound

$\|\delta(s, c)\|_\infty \le \sigma(g_{max})$



For every input s and every c, since each factor has magnitude below 1.

Theorem 2: monotonicity

$\operatorname{sign} \partial\delta_j/\partial c = \operatorname{sign} u_j(s)$



Each coordinate moves one way in c, set by the input. Rising refusal is observed, not proved.

Lemma 1: Lipschitz in c

$\|\delta(c) - \delta(c')\|_\infty \le \sigma(g_{max}) \lvert c - c' \rvert / T$

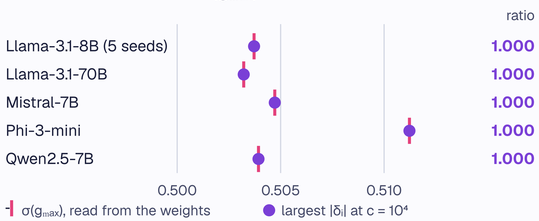


A unit change in c moves the residual by at most 0.10 on Llama-3.1-8B.

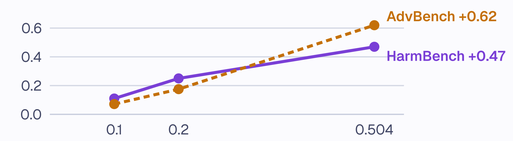
Training: 5,000 HH-RLHF preference pairs, the chosen reply at c = +1 and the rejected reply at c = −1, with LoRA and the DIAL head trained jointly.

3 Bound in practice

The residual reaches $\sigma(g_{max})$ on five base models.



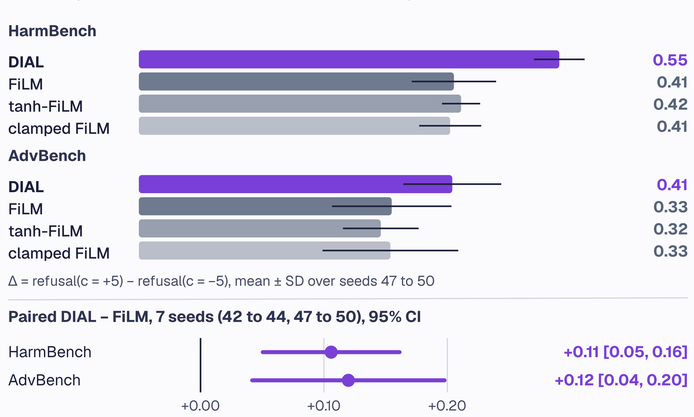
Clamping the gate in training sets the bound and narrows control



Δ = refusal(c = +5) − refusal(c = −5) with $\sigma(g_{max})$ clamped to 0.1 or 0.2 in training, or left free at 0.504; seed 42.

4 Matched FiLM variants

Capping FiLM's residual does not close the gap.



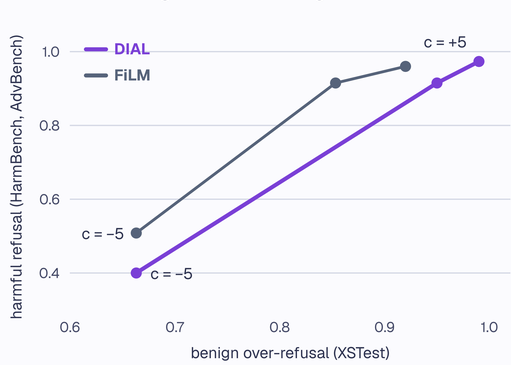
Δ = refusal(c = +5) − refusal(c = −5), mean ± SD over seeds 47 to 50

Paired DIAL − FiLM, 7 seeds (42 to 44, 47 to 50), 95% CI

Same objective and data for all four heads. tanh-FiLM applies tanh to each residual coordinate; clamped FiLM clips each at ±0.5.

5 Refusal trade-off

Harmful and benign refusal rise together.



XSTest values pending a re-run on matched prompts.

Mean of seeds 42 to 44, points at c = −5, 0, +5. GPT-4o judge, κ = 0.895 against 50 human labels.

6 Scope

The bound applies to the residual, not to behaviour.

	bounded	monotone	scalar c	coherent
DIAL	✓	✓	✓	✓
FiLM	✗	|c| ≤ 5	✓	✓
CAA	✗	✗	✓	✓
TempScale	✗	✗	✓	✗
LoRA	✗	n/a	✗	✓

TempScale multiplies the logits by 1 + c/T and produces gibberish at c ≤ −5.

Outside the bound

- Logits, token probabilities and refusal rates.
- The L_2 norm: at saturation it is 27 to 44 across the five models.
- Inferred control values: a meta-RL policy that infers c from its history trails a plain MLP on 2 of 3 seeds.

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.