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

Nguyen Vu Nguyen
University of Wollongong
nvn994@uowmail.edu.au

Abstract

We introduce DIAL (Directional Intervention, Asymptotically Limited), a parameter-efficient adapter that adds a scalar-controlled residual to a base language model's hidden state. The residual is the elementwise product of a learned per-coordinate sigmoid gate $\sigma(g)$, a saturating function $\tanh(c/T)$ of the control value c, and a direction computed from the hidden state and bounded in $[-1, 1]$ per coordinate. Because the last two factors never exceed 1 in magnitude, the residual L_∞ norm is bounded by the largest gate value $\sigma(g_{\max})$, $g_{\max} = \max_i g_i$, for every input and every c, and the bound is read from the weights alone. The bound concerns the hidden-state residual and does not extend to token probabilities or refusal behavior. Across 5 base models (3.8B to 70B parameters, 4 architecture families, one checkpoint per model), the ratio of the realized residual norm at $|c| = 10^4$ to the bound is 1.000 on all five models, an empirical tightness result separate from the validity of the bound. On Llama-3.1-8B under a GPT-4o judge ($n = 100$ prompts per benchmark), the controllability range, the change in refusal rate from $c = -5$ to $c = +5$, exceeds that of a feature-wise linear modulation (FiLM) baseline trained with the same objective: HarmBench $+0.52$ vs $+0.44$ and AdvBench $+0.63$ vs $+0.46$ (means over three paired seeds), with a pooled advantage of $+0.11$ (95% CI ± 0.06) over seven paired seeds, and neither a tanh-controlled FiLM nor a FiLM with its residual clipped to 0.5 closes the gap. Benign over-refusal on XSTest rises with c at comparable rates for both methods ($+0.33$ vs $+0.26$), and DIAL does not dominate the safety-helpfulness trade-off. At $c = 0$ the residual is exactly zero and the adapted model reproduces the base model. Outside the training range DIAL's residual stays at the bound while FiLM's grows linearly, reaching $534\times$ DIAL's at $|c| = 10^4$ on a fixed Gaussian probe input. The trained gate bound is about 0.5, and clamping it lower narrows the controllability range. Code, adapter checkpoints, model generations for all benchmarks, the GPT-4o judge cache, and a human-labeled validation set ($n = 50$, agreement with the judge $\kappa = 0.895$) are released.

1 Introduction

Adapters, steering vectors, and preference optimization all modify a model's residual stream, but none gives a closed-form magnitude bound on that modification. This matters for safety claims. They can be trained, evaluated, measured empirically; they do not admit a single number documenting how much the representation can change as a function of the control parameter. Post-hoc fits estimate this from data; they do not bound it from weights. For FiLM, prefix tuning, CAA, and RepE, the worst-case perturbation requires adversarial estimation across c; LoRA, RLHF, and DPO admit no scalar control at all.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A *structural prior* on the residual form solves this directly. We study

$$\delta(s, c) \;=\; \sigma(g) \odot \tanh(c/T) \odot \tanh\big(u(s)\big), \quad u(s) \;:=\; W\phi(s) + b, \tag{1}$$

where σ is the sigmoid, $g \in \mathbb{R}^d$ a learned per-dimension gate, T a temperature, $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ the direction layer, $\phi(s) \in \mathbb{R}^d$ the final hidden state, and $u(s)$ the direction pre-activation. We call this *DIAL*: *D*irectional *I*ntervention, *A*symptotically *L*imited.

The argument is structural: (1) gives a closed-form residual L_∞ bound and sign-monotone scalar control directly from the algebra. The residual induced by any control value c is bounded by $\sigma(g_{\max})$, computable from weights alone. Among compared methods, only the bounded-monotone form satisfies bound, monotonicity, and Lipschitz scaling simultaneously: FiLM has monotonicity in-distribution but no bound; CAA has neither (non-monotone between training endpoints, Section 7); TempScale (an unbounded variant of our method ablated in Section 7) loses capability at $|c| \geq 5$; LoRA and DPO have no scalar control. We provide evidence at three model scales (3.8B–70B, 4 architectures) and as a second-domain check on continuous-control RL (Appendix D); diffusion is treated as algebraic observation only.

Contributions.

- **Structural properties.** The residual $\delta = \sigma(g) \odot \tanh(c/T) \odot \tanh(u(s))$ yields three properties without any training-time assumptions: a closed-form L_∞ bound (Theorem 1), sign-monotone scalar control (Theorem 2), and c-Lipschitz scaling (Lemma 1). No other compared method satisfies all three.
- **Empirical tightness of the bound.** The bound holds by construction and requires no measurement; on the evaluated prompts the realized residual reaches it, with ratio $\|\delta\|_\infty/\sigma(g_{\max}) = 1.000$ on all 5 base models and on each of 5 Llama-8B seeds, verified across 9 ablations. (Section 5.)
- **Behavioral comparison.** Under GPT-4o (n=100 per bench), DIAL's refusal controllability range on HarmBench and AdvBench exceeds FiLM's under both semantic and strict-phrase judges, and exceeds that of tanh-controlled and clamped FiLM variants over four matched seeds. Benign over-refusal (XSTest) is a comparable cost for both methods. (Section 6.)
- **Architectural bound at OOD c.** The closed-form $\|\delta\|_\infty \leq \sigma(g_{\max})$ holds at all c, computable from weights alone; FiLM's residual reaches $534\times$ DIAL's at $|c|$=10^4. (Section 6, Table 5.)

Deployment relevance. The bound matters beyond adversarial extremes. *Configuration error*: a unit bug producing c=50 instead of c=5 leaves DIAL bounded at $\sigma(g_{\max})$, while FiLM's residual grows $\sim 5\times$. *Multi-tenant deployment*: with c exposed to per-customer configuration, DIAL's bound is enforced by the weights, while FiLM's depends on every consumer correctly clamping c. *Adversarial composition*: an attacker who sets c arbitrarily leaves DIAL at $\sigma(g_{\max})$ but admits unbounded FiLM residual injection. An architectural bound shifts the safety property from deployment code, which every consumer must maintain, into model weights, which the model owner controls. The analogy is memory safety: a typed language prevents classes of bugs that scattered runtime checks can only catch case by case. Section 6 (Table 5) and Figure 3 make the contrast empirical.

Scope. We claim a closed-form L_∞ residual bound computable from weights alone, monotone scalar controllability, and refusal controllability on HarmBench and AdvBench that exceeds FiLM's under both GPT-4o judges (n=100). Other compared methods (CAA, TempScale, LoRA, DPO+DIAL) either lack controllability or break monotonicity. We also release the full reproducibility pipeline. We do *not* claim the dial creates safety from scratch on non-aligned models; that capability is fully preserved (MMLU shows a 0.5–1.5pp gap); that the residual bound implies formal guarantees on token probabilities, decoded trajectories, or refusal behavior (monotone refusal in c is an empirical finding); that DIAL dominates FiLM on the safety-helpfulness trade-off; or that the prior generalizes to non-refusal axes or to settings where c must be inferred online. The RL second-domain check (Appendix D) does not constitute broad cross-domain generalization.

2 Related work

Steering vectors and activation editing. CAA [29] and activation addition [38] add a fixed direction v to hidden states with scalar strength c: $h' = h + c \cdot v$. More sophisticated variants exist:

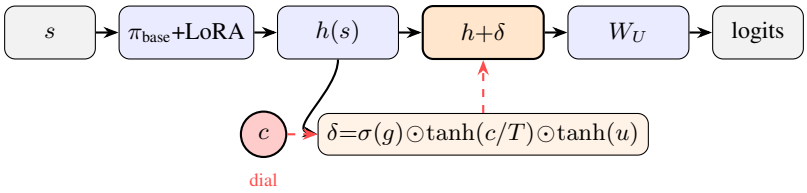

Figure 1: DIAL architecture. Hidden state $h(s)$ from base+LoRA is augmented by the DIAL residual δ; the control dial c enters via $\tanh(c/T)$, providing the asymptotic bound of Theorem 1.

RepE [42] applies linear combinations of representation directions across layers; ITI [21] scales interventions by per-head standard deviations across selected attention heads; weight-arithmetic steering [13] operates in parameter space. Despite their differences, all inject an intervention that grows unboundedly in $|c|$ and is empirically non-monotone outside a narrow range. Activation patching [27, 8] edits hidden states for interpretability but does not impose structural bounds. Our negative result on CAA (Section 7) shows the U-shape failure mode of unbounded steering. Constitutional classifier work [35] addresses related defense concerns but without formal bounds on the residual norm. Sparse autoencoders [37] decompose activations into human-readable feature directions, cleaner semantically than DIAL's learned D but without a closed-form magnitude bound, monotonicity guarantee, or single-scalar certification. DIAL trades interpretability for those three structural properties; the two approaches address different needs.

Adapters and PEFT. LoRA [17], QLoRA [11], FiLM [30], prefix tuning [22], and P-tuning [25] attach learned residuals. None provides a closed-form residual bound; FiLM-with-conditioning gives scalar control via its γ, β functions but the resulting residual grows unboundedly in the conditioning input (we use FiLM as our primary c-controllable baseline, Section 6); LoRA, prefix tuning, and P-tuning in their standard formulations have no scalar control parameter at all. We show standard LoRA achieves zero refusal-controllability range under matched LLM-judged evaluation. The DIAL prior is composable with LoRA (we apply both jointly).

Structural priors. Monotonic networks [10] and Lipschitz-controlled architectures [2] share the principle that structural constraints provide guarantees empirical fits cannot. DIAL applies this principle in a narrow setting: a closed-form magnitude bound on a scalar-conditioned adapter residual, rather than a representation-level structural commitment. Our contribution is the specific architectural form that makes $\sigma(g_{\max})$ a weight-readable upper bound; the structural-prior principle itself is older.

RLHF, DPO, and constitutional methods. RLHF [36, 28, 3] and DPO [32] produce non-controllable models (no scalar c to dial). Constitutional AI [4] adds rule-following supervision but again without continuous control. Our negative result N4 (Section 7) shows that naive DPO+DIAL composition fails; composability requires joint training.

Diffusion guidance and CFG. Classifier guidance [12] and classifier-free guidance [16] on Stable Diffusion [33] and SDXL [31] is the diffusion analogue of activation steering: $\epsilon_w = \epsilon_u + w(\epsilon_c - \epsilon_u)$, unbounded in w. The same algebraic form as DIAL admits an analogous bound for guidance-scale modulation; behavioral evaluation in diffusion is left to future work.

Safety benchmarks and judged evaluation. HarmBench [26], AdvBench [43], XSTest [34], BeaverTails [19], JailbreakBench [5], SORRY-Bench [39] are the standard safety benchmarks. Capability evaluation uses MMLU [15], GSM8K [7], TruthfulQA [23], HumanEval [6], IFEval [41]. We use LLM-as-judge [40] (GPT-4o) following best practice; we demonstrate regex-based scoring is unreliable for this task concretely in Appendix D (regex yields $0.00/0.00/0.02$ where the judge yields $0.57/0.84/0.96$ on the same 70B generations).

3 Method: DIAL

Why this form? Three bounded factors (a per-dimension gate $\sigma(g)$, a saturating control $\tanh(c/T)$, and a bounded direction $\tanh(u(s))$) combine to give the bound, sign monotonicity, and Lipschitz scaling of Section 4 simultaneously. Removing the bounded factor (CAA) breaks monotonicity

(Section 7); removing the saturating-c factor (TempScale) breaks capability preservation. Applying $\tanh(c/T)$ to FiLM's scale parameter does not yield a closed-form weight-only bound: bounding γ via $\tanh$ alone leaves the residual $\gamma h + \beta$ unbounded (since $\beta(c)$ grows in c), and bounding β as well only gives $\|h\|_\infty + 1$, an input-dependent quantity. DIAL avoids the multiplicative $\gamma \odot h$ structure entirely, replacing it with $\sigma(g) \odot \tanh(u(s))$ where both factors are bounded transformations whose product yields a constant $\sigma(g_{\max})$ computable from a single weight vector, independent of input. A weight-only, input-independent bound is the deployment-relevant property; FiLM's form does not admit one under any $\tanh$ wrapping that preserves its conditional structure. Behaviorally, tanh-controlled and clamped FiLM heads trained with the same objective also fall short of DIAL's controllability range (Table 4).

Architecture. Given a base model π_{base} with deep hidden state $h(s) \in \mathbb{R}^d$, DIAL adds the residual (1) to the final hidden state before the unembedding W_U:

$$\pi_{\text{DIAL}}(s, c) \;=\; W_U \big[\, h(s) + \sigma(g) \odot \tanh(c/T) \odot \tanh\big(W h(s) + b\big) \,\big]. \tag{2}$$

Components:

- $g \in \mathbb{R}^d$: a learned per-coordinate gate, initialized to zero so $\sigma(g) = 0.5$ at start, and trained without a target value.
- $T = 5$: fixed temperature; ablations at $T \in \{1, 15, 50\}$ appear in Table 2 (T=15), Appendix A.3 (Lipschitz slope), and Appendix E (T=1 and T=50 in multi-turn and injection tables).
- $W \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$: direction layer, $W \sim \mathcal{N}(0, 0.02^2)$, $b = 0$.
- $\phi(s) := h(s) \in \mathbb{R}^d$: state encoding tied to the base hidden state (parameter-free).
- LoRA [17] of rank 8 on all 7 base projection modules (attention q, k, v, o and MLP gate, up, down) is applied jointly. The DIAL head and LoRA are trained together end-to-end.

Total trainable: $\sim$ 37.8M parameters for Llama-3.1-8B (**0.47%** of base 8.0B; DIAL head $\approx$ 16.8M + LoRA $\approx$ 21.0M); $\sim$ 170.7M for Llama-3.1-70B (**0.24%** of base). Appendix B works through the residual on a trained checkpoint.

Training objective. We train on preference pairs (x, y^+, y^-) from Anthropic HH-RLHF [3], where y^+ is the preferred and y^- the rejected continuation, with a dual-c loss:

$$\mathcal{L}(\theta) \;=\; \tfrac{1}{2}\, \mathbb{E}_{(x,y^+,y^-)} \big[\mathcal{L}_{\text{LM}}(\pi_\theta(\cdot, c{=}{+}1),\, x \oplus y^+) + \mathcal{L}_{\text{LM}}(\pi_\theta(\cdot, c{=}{-}1),\, x \oplus y^-) \big], \tag{3}$$

where $\mathcal{L}_{\text{LM}}$ is next-token cross-entropy over the full conversation. The sign of c is the only difference between the two terms, and the learned direction separates preferred from rejected behavior. A further DIAL variant trained on UltraFeedback [9] prompts serves as a dataset-ablation arm, denoted "UltraFeedback" in extension tables.

Hyperparameters (default). 5k steps with one preference pair per step (the first 5k pairs of a 16k-example HH-RLHF subset), AdamW, learning rate $1\text{e}{-4}$, weight decay 0.01, gradient-norm clipping 1.0, maximum length 384 tokens, BF16 compute, 4-bit NF4 base quantization for $\geq$ 7B models. Inference: $c \in \mathbb{R}$ (no train-time clamp); empirical evaluation uses $c \in [-10, 10]$ for behavior, $c \in [0, 10^4]$ for bound saturation.

4 Theory

Theorem 1 (Residual saturation, L_∞ form)**.** *For any input s and any $c \in \mathbb{R}$,*

$$\big\|\delta(s, c)\big\|_\infty \;\leq\; \sigma(g_{\max}) \cdot |\tanh(c/T)|, \tag{4}$$

where $g_{\max} = \max_i g_i$. The bound is asymptotically attained on inputs satisfying the saturation condition $|\tanh(u_{i^*}(s^*))| \to 1$ for $i^* = \arg\max_i \sigma(g_i)$: for such s^*, $\|\delta(s^*, c)\|_\infty \to \sigma(g_{\max})$ as $|c| \to \infty$. We verify this empirically across all 5 base models (Section 5); the bound itself holds by algebra independent of input distribution.

Proof sketch. Each output coordinate i satisfies $|\delta_i| = |\sigma(g_i) \cdot \tanh(c/T) \cdot \tanh(u_i(s))| \leq \sigma(g_i) \cdot |\tanh(c/T)|$, where the last factor uses $|\tanh(\cdot)| \leq 1$. Taking $\max_i$ gives (4). Saturation: choose any input where $|\tanh(u_i(s))| \geq 1 - \epsilon$ for the argmax-g_i coordinate; full proof in Appendix A. $\square$

Table 1: Empirical tightness of Theorem 1 (csweep on held-out prompts at $|c| = 10^4$). The bound holds by construction; r measures how closely each checkpoint reaches it. "$n = 5$" refers to 5 csweep seeds (s42–s46); the $n = 3$ safety-bench metric in Section 6 is distinct.

Family	$\sigma(g_{\max})$	$\|\delta\|_\infty(c{=}10000)$	Ratio r
DIAL-Llama-8B (n=5)	0.5037	0.5037	1.000
DIAL-Llama-70B	0.5032	0.5032	1.000
DIAL-Mistral-7B-Instruct-v0.3	0.5047	0.5047	1.000
DIAL-Phi-3-mini-4k	0.5112	0.5112	1.000
DIAL-Qwen2.5-7B-Instruct	0.5039	0.5039	1.000
Mean	≈ 0.505	≈ 0.505	**1.000**

Theorem 1 makes two statements of different status. The inequality (4) holds for every input and every c by construction, independent of training and data. Tightness, the realized residual reaching $\sigma(g_{\max})$, depends on whether the trained direction layer drives inputs into saturation and is an empirical property of each checkpoint on the evaluated prompts (Table 1: ratio 1.000 on all 5 base models). The bound concerns the hidden-state residual only. It does not bound token probabilities, decoded trajectories, or the probability of harmful compliance, and an L_∞ norm near 0.5 admits an L_2 norm up to $\sqrt{d}\,\sigma(g_{\max}) \approx 32$ at $d{=}4096$, comparable to typical hidden-state norms.

Theorem 2 (Sign monotonicity)**.** *For each output coordinate i and any input s, the function $c \mapsto \delta_i(s, c)$ is monotone on $\mathbb{R}$, with the direction of monotonicity given by* $\operatorname{sign}(u_i(s))$*:*

$$\operatorname{sign}(\partial\delta_i/\partial c) \;=\; \operatorname{sign}\big(u_i(s)\big) \quad \textit{independently of } c.$$

Lemma 1 (c-Lipschitz)**.** $\|\delta(s, c) - \delta(s, c')\|_\infty \le \frac{\sigma(g_{\max})}{T}|c - c'|$.

Full proofs in Appendix A. Theorems 1–2 and Lemma 1 are structural results holding by architectural construction. Because $\delta(s, c) = \tanh(c/T)\, v(s)$ with $v(s) := \sigma(g) \odot \tanh(u(s))$ independent of c, every linear functional of the residual, including each next-token logit $W_U[h(s) + \delta(s, c)]$ for a fixed prefix, is monotone in c. Probabilities after the softmax, multi-token decoding, and refusal rates are not covered; the monotone refusal curves of Section 6 are empirical. Capability preservation is empirical, not architectural, and is verified separately in Section 6 and Appendix D.

5 Empirical verification: saturation ratio

We verify Theorem 1 empirically across 5 base models (one checkpoint each) with 5-seed variance at Llama-3.1-8B, additionally covering 9 Llama-8B ablations. For each variant, we compute $\|\delta(s, c)\|_\infty$ on a held-out prompt set across $c \in \{0, 1, 2, 5, 10, 100, 1000, 10000\}$. Figure 2 shows the saturation curves; Table 1 reports the saturation ratio $r := \|\delta\|_\infty(c{=}10000)/\sigma(g_{\max})$. At $|c| \ge 100$, 100% of evaluated (seed, prompt, c) triples achieve saturation ratio > 0.99; the bound is not loose for any evaluated input. Four further Llama-3.1-8B seeds (s47–s50) give $\sigma(g_{\max}) = 0.5036 \pm 0.0003$, with the residual reaching the bound on every seed. We additionally verify saturation behavior on continuous-control RL policies (HC-Vel $n{=}35$, Ant-Goal $n{=}30$, oracle-c; Appendix D); the prior fails when c is inferred online (Appendix J).

6 Safety controllability at 8B

We compare DIAL against 6 structural alternatives on 3 safety benchmarks (HarmBench, AdvBench, XSTest) at the Llama-3.1-8B scale. Generations are scored by GPT-4o as judge for refusal classification (REFUSE / COMPLY / PARTIAL).

Judge selection. We use GPT-4o as the canonical judge throughout this section, evaluated on $n{=}100$ prompts per benchmark with zero Azure content-filter blocks observed, under both semantic and strict-phrase refusal-detection prompts. We selected GPT-4o based on capability and the absence of content-filter blocks at our evaluation scale, which avoids conditioning Δ on filter-passing samples. Validation against author labels yielded 94% agreement, Cohen's $\kappa{=}0.895$ (Judge validation paragraph below).

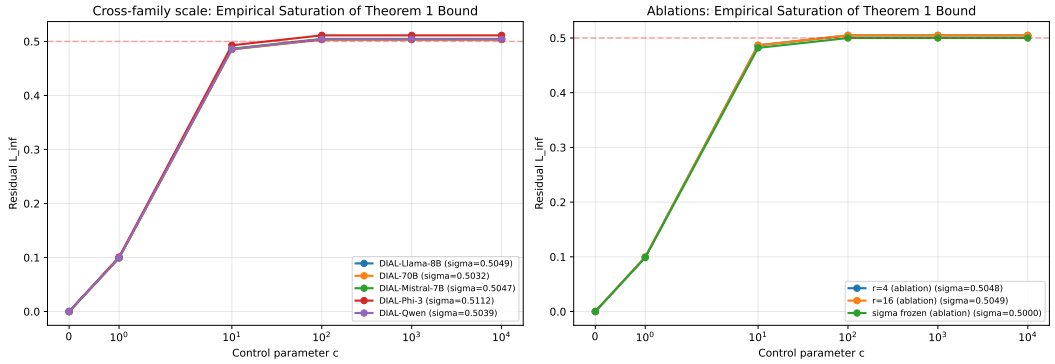

Figure 2: **Empirical saturation of the residual bound (Theorem 1).** Left: across model families (8B–70B). Right: ablations. Every DIAL variant saturates at $\sigma(g_{\max})$ as $|c| \to \infty$.

Table 2: Behavioral controllability range $\Delta = \text{refuse}(c{=}5) - \text{refuse}(c{=}{-}5)$ at Llama-3.1-8B under GPT-4o judge (n=100 per bench, zero content-filter blocks). Format: *semantic / strict*. XSTest Δ is benign over-refusal, a cost. Aggregates from `analysis_results/judge_results_gpt4o.json`.

Arm	n	Δ(HB)	Δ(AdvBench)	Δ(XSTest)†	MMLU
Behavioral controllability:					
DIAL	3	$\mathbf{+0.520}$ / $\mathbf{+0.550}$	$\mathbf{+0.627}$ / $\mathbf{+0.657}$	$\mathbf{+0.327}$ / $\mathbf{+0.250}$	0.615
FiLM	3	$+0.443$ / $+0.393$	$+0.460$ / $+0.500$	$+0.257$ / $+0.223$	0.620
Limited or no controllability (semantic-judge Δ):					
σ-frozen (DIAL ablation)	1	$+0.290$	$+0.130$	$+0.120$	0.610
T=15 (DIAL ablation)	1	$+0.290$	$+0.290$	$+0.250$	0.600
CAA	1	$+0.170$	$+0.130$	$+0.260$	0.620
TempScale‡	2	-0.070	-0.030	-0.122	0.620
LoRA (no c-knob)	1	0.000	0.000	0.000	0.630
DPO+DIAL hybrid (degen.)	1	$+0.010$	$+0.020$	-0.020	0.620

†XSTest $\Delta > 0$ means the dial increases over-refusal on benign prompts (safety-helpfulness tradeoff), not safety improvement.

‡TempScale's c<0 output is gibberish (logit-zeroing); $\Delta < 0$ reflects this. TempScale s42 XSTest was evaluated on n=90 (10 content-filtered); all other cells n=100.

Judge validation. We validated the GPT-4o judge against author labels on n=50 generations from the judge cache (multiple arms, benchmarks, c values): 94% agreement, Cohen's $\kappa = 0.895$ [20]. All 3 disagreements were PARTIAL↔COMPLY; **zero** REFUSE↔COMPLY errors (the only kind affecting controllability-range claims). Released as `analysis_results/human_eval.csv`.

Judge precedent and regex failure. LLM-based safety classifiers are standard: Llama Guard [18], HarmBench's classifier [26], JailbreakBench's judge [5]. Regex is not a viable alternative: on identical 70B generations, regex yields $0.00/0.00/0.02$ where the judge yields $0.57/0.84/0.96$, consistent with [34] on keyword matching missing avoidance-style refusals.

Headline metric: controllability range. For each arm, we report $\Delta := \text{refuse}(c{=}5) - \text{refuse}(c{=}{-}5)$. A controllable arm has Δ approaching the safety-prior trained range; a flat (uncontrollable) arm has $\Delta \approx 0$. On HarmBench and AdvBench Δ measures control of harmful refusal; on XSTest, whose prompts are benign, Δ measures the over-refusal cost of raising c.

Behavioral comparison. On HarmBench and AdvBench DIAL's controllability range exceeds FiLM's under both judges (Table 2). Over the three paired seeds the semantic-judge Δ is 0.52 ± 0.05 for DIAL against 0.44 ± 0.06 for FiLM on HarmBench and 0.63 ± 0.01 against 0.46 ± 0.07 on AdvBench (mean and SD over seeds; paired DIAL$-$FiLM 0.077 ± 0.040 and 0.167 ± 0.068). Over seven paired seeds, the three above and four further matched seeds (s47–s50), Δ is 0.54 against 0.43 on HarmBench (paired $+0.11 \pm 0.06$) and 0.50 against 0.38 on AdvBench (paired $+0.12 \pm 0.08$), and the pooled harmful-bench paired advantage is $+0.11$ (95% CI ± 0.06), which excludes zero. The gap is modest, and the contribution rests on the bound (Table 5). On the four further seeds

Table 3: Harmful refusal (mean of HarmBench and AdvBench) and benign over-refusal (XSTest) against c, GPT-4o semantic judge, n=100 per bench, mean over three paired seeds. DIAL at c=0 is the base model.

c	DIAL harmful	DIAL over-refusal	FiLM harmful	FiLM over-refusal
-5	0.400	0.663	0.508	0.663
0	0.915	0.950	0.915	0.853
$+5$	0.973	0.990	0.960	0.920

Table 4: Controllability range $\Delta = \text{refuse}(c{=}5) - \text{refuse}(c{=}{-}5)$ under the GPT-4o semantic judge (n=100 per bench), mean $\pm$ SD over four matched seeds (s47–s50). XSTest Δ is benign over-refusal, a cost.

Method	HarmBench	AdvBench	XSTest
DIAL	$\mathbf{0.55 \pm 0.03}$	$\mathbf{0.41 \pm 0.06}$	0.29 ± 0.02
FiLM	0.41 ± 0.06	0.33 ± 0.08	0.24 ± 0.04
tanh-FiLM	0.42 ± 0.02	0.32 ± 0.05	0.23 ± 0.04
clamped-FiLM	0.41 ± 0.04	0.33 ± 0.09	0.27 ± 0.05

AdvBench Δ is 0.41 because refusal at $c{=}{-}5$ starts higher (0.555 against 0.360), while refusal at $c{=}{+}5$ reproduces (0.965 against 0.987). At $c{=}{+}5$ both methods reach similar refusal rates, and DIAL is more compliant at $c{=}{-}5$. Other compared methods are worse: TempScale's $\Delta < 0$ reflects logit-zeroing degeneration at $c{<}0$; CAA shows reduced controllability and is non-monotone in c (Section 7); LoRA produces $\Delta = 0$ by construction; the DPO+DIAL hybrid produces near-zero Δ across all benches (N4, Section 7).

Safety-helpfulness trade-off. Table 3 sets harmful refusal (mean of HarmBench and AdvBench) against benign over-refusal (XSTest). At c=0, $\delta(s, 0) = 0$ and DIAL reproduces the base model, whose benign refusal is 0.950; FiLM at c=0 already departs from the base model. DIAL's residual is a single direction scaled by $\tanh(c/T)$, so harmful and benign refusal move together and DIAL cannot lower one without the other. FiLM's per-dimension modulation decouples them, at the cost of the base anchor, the bound, and monotonicity. On this trade-off the two methods are comparable and neither dominates. Separating benign from harmful control while keeping a weight-readable bound, for example with several bounded directions, is left to future work.

Bounded FiLM variants. A tanh-controlled FiLM and a clamped FiLM, whose residual is clipped to the constant 0.5 to match DIAL's $\sigma(g_{\max})$, were trained with the same objective and judged in the same pipeline over four seeds (Table 4). DIAL's harmful-bench range is the largest; the unconstrained, tanh, and clamped FiLM heads cluster below it. The clamped head has the same worst-case residual budget as DIAL, so the difference is not one of allowed magnitude. On XSTest the four methods are comparable. Neither variant has a weight-readable bound: the tanh-FiLM residual caps at the constant 1 regardless of its weights while its pre-tanh modulation reaches $L_\infty \approx 2050$ at $|c|{=}10^4$, and the clamp is a fixed constant rather than a per-coordinate value read from trained weights.

Cross-judge robustness. Per-arm cross-judge analysis on the n=30 subsample (same prompts for both judges; note the Table 2 values are from the full n=100 run) and multi-turn / injection / long-context stress tests in Appendix F. The architectural bound $\sigma(g_{\max})$ is judge-independent.

Bound comparison. The behavioral comparison above is in-distribution; the residual bound is what separates DIAL from FiLM *out-of-distribution*. Table 5 and Figure 3 show DIAL's residual saturates at $\sigma(g_{\max}) \approx 0.50$ for $|c| \geq 100$ by architectural construction (Theorem 1), while FiLM's grows linearly to 534× DIAL's at $|c|{=}10^4$. At $|c| \leq 5$ FiLM's residual is smaller than DIAL's; the bound advantage manifests at OOD c.

The bound is non-trivial precisely because it must hold without input-distribution assumptions. Empirical bounds estimated from training data (e.g., percentile bounds on observed residual norms) provide no guarantee outside the training distribution; certified bounds via Lipschitz constraints on the full network introduce capability costs that DIAL avoids by localizing the constraint to the residual; weight-space norms (e.g., $\|D\|_2$) bound a different quantity than the residual itself. DIAL's bound

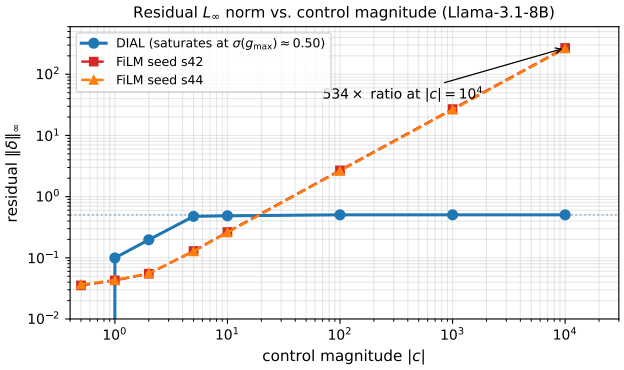

Figure 3: Residual L_∞ norm vs. $|c|$ (log-log, Llama-3.1-8B). DIAL saturates at $\sigma(g_{\max}) \approx 0.504$ for $|c| \geq 100$; FiLM grows linearly to ≈ 269 at $|c|{=}10^4$ ($534\times$ DIAL). At $|c|{\leq}5$ FiLM is smaller; the bound advantage is at OOD c.

Table 5: Residual L_∞ norm at increasing $|c|$. DIAL saturates by architectural construction; FiLM grows linearly without bound. DIAL data are batch-averaged empirical residuals from the `ood_bound_check` field of canonical csweeps; FiLM values apply each trained FiLM head to a fixed Gaussian probe with $\|h\|_2 \approx 32$; the residual is affine in c for any probe, and two independently trained FiLM seeds give 269 and 264 at $|c| = 10^4$. The exact multiple depends on the probe.

$\lvert c\rvert$	DIAL $\lVert\delta\rVert_\infty$	FiLM s42 $\lVert\delta\rVert_\infty$	FiLM s44 $\lVert\delta\rVert_\infty$	FiLM/DIAL ratio
0	0.00	0.04	0.04	—
1	0.10	0.04	0.04	$0.4\times$ (FiLM smaller)
5	0.48	0.13	0.13	$0.3\times$ (FiLM smaller)
10	0.49	0.26	0.26	$0.5\times$ (FiLM smaller)
100	**0.50** (sat.)	2.69	2.63	$5.4\times$
1000	**0.50** (sat.)	26.94	26.35	$53\times$
10000	**0.50** (sat.)	269.4	263.5	$\mathbf{534}\times$

$\sigma(g_{\max})$ is computable from a single weight vector, holds at all c, and is tight up to input-distribution coverage of the saturating regime.

Capability note. DIAL's MMLU (0.615) is 0.5pp below FiLM (0.620) and 1.5pp below LoRA (0.630). Reported seed variance for 8B-class instruction-tuned models on MMLU is typically $\pm$0.5–1.0pp, so the FiLM gap is plausibly within noise. The LoRA gap is not from residual perturbation at evaluation ($\delta(s, 0) = 0$ exactly at $c{=}0$); the likelier cause is the multi-c training objective shifting the $c{=}0$ optimum relative to LoRA's single-target optimization. Direct seed-variance and broader capability evaluation are left to future work.

7 Ablations and structural failure modes

We collect ablations that violate components of the DIAL form. Two are structural failure modes (N1, N2) showing that bound and saturating temperature are each necessary; one is a parsimonious bounded design point (N3) that retains the closed-form bound at lower controllability magnitude; one is a pipeline-composition failure (N4); and one varies the gate bound (N5).

(N1) CAA (unbounded steering). Removing the $\sigma(g) \odot \tanh$ envelope produces non-monotone behavior. The csweep refusal trace (`csweeps/caa_s42_csweep.json`, evaluated on the multi-c csweep harmful prompts ($n{=}100$) rather than the HarmBench bench set used in Table 2) shows a U-shape across $c \in [-10, 10]$: refusal $0.79, 0.95, 0.47, 0.44, 0.42, 0.48, 0.45, 0.49, 0.85$ at $c = -10, -5, -2, -1, 0, 1, 2, 5, 10$, violating Theorem 2. The csweep prompt set and HarmBench prompt set are distinct; both legitimately measure CAA behavior, and the csweep trace is where the U-shape (non-monotonicity in c) is visible. Structural boundedness is what enables monotonicity.

(N2) TempScale (no saturating-c factor). Replacing $\tanh(c/T)$ with c degenerates the model to gibberish at $c \leq -5$, where the logit scale $1 + c/T$ reaches zero and then turns negative, violating capability preservation.

(N3) σ-frozen, a parsimonious bounded variant. We frame N3 as a design-point variant rather than a failure: fixing the gate g at initialization (only D trained) preserves Theorem 1. Under GPT-4o, σ-frozen achieves HarmBench $\Delta = +0.290$, AdvBench $+0.130$, XSTest $+0.120$, all below DIAL's full-trainable values but retaining the closed-form residual bound that FiLM lacks. The right design point depends on deployment requirements: σ-frozen for parsimony plus bound; DIAL for higher controllability plus bound; FiLM for higher controllability without bound.

(N4) 2-phase DPO+DIAL hybrid. We trained DIAL jointly with SFT and then applied DPO. The training loss converges (SFT $2.774 \to 1.702$, DPO $0.694 \to 0.305$), but at inference the model emits non-English degenerate output (Russian *"zreniya" (vision)* loops) for all c, and the dial has near-zero behavioral effect ($\Delta \in \{+0.010, +0.020, -0.020\}$ across HB/AdvBench/XSTest under GPT-4o). The collapse is consistent with reward over-optimization pathologies documented for preference-learning objectives [24, 14]. We ran a post-hoc spectral analysis of the direction weight W in both the base DIAL checkpoint and the DPO+DIAL checkpoint. Comparing W after DPO against W before DPO, the leading singular value drops by 30% ($9.625 \to 6.736$), σ_2 shifts modestly ($4.185 \to 4.582$, $\sim 10\%$), and higher-order singular values remain stable ($< 1\%$ change). The leading singular value ratio σ_1/σ_2 collapses from $2.30\times$ to $1.47\times$, indicating that DPO suppresses the primary safety-relevant direction. We call this *principal component diffusion*: the safety axis becomes no more prominent than noise directions, even though the full rank is preserved. This is consistent with the behavioral collapse: the dial remains architecturally present, but the dominant direction it modulates is suppressed after DPO (rank is preserved). We did not test this mechanism by intervention. We do not claim the DIAL structural prior is incompatible with DPO in principle; we report that *naive sequential composition*, DPO applied after DIAL training, fails.

(N5) Gate bound as a design parameter. The gate is initialized at zero, so $\sigma(g) = 0.5$ at start, and is trained without a target; $\sigma(g_{\max}) \approx 0.5$ is not tuned. Clamping the gate to 0.2 and 0.1 and retraining seed 42 with the same objective and judge holds the realized bound at exactly 0.200 and 0.100. The controllability range narrows monotonically as the bound tightens: HarmBench $\Delta = +0.47, +0.25, +0.11$ at bound $0.5, 0.2, 0.1$ (AdvBench $+0.62, +0.18, +0.07$; XSTest $+0.39, +0.13, +0.10$), where the 0.5 row is the submitted seed-42 model. The narrowing is on the $c < 0$ side: refusal at $c{=}{+5}$ stays between 0.92 and 1.00 on HarmBench, while at $c{=}{-5}$ the smaller bound remains nearer the base model (0.89 at bound 0.1 against 0.45 at 0.5), because a smaller cap permits less deviation from the base model in either direction. The bound trades worst-case residual magnitude against controllability range, and 0.5 is one operating point on that trade-off. These are single-seed runs.

The negative finding is twofold: practitioners should not stack alignment objectives without joint optimization; any successful hybrid will require single-phase joint training, which we leave to future work. The DIAL configuration sits at one end of a small bounded-monotone family; we view the family, not any single member, as the contribution.

8 Discussion and limitations

What is the right family? DIAL is one tractable member of the bounded-monotone family. Per-property accounting: DIAL and σ-frozen yield the bound; FiLM lacks it; CAA lacks bound and monotonicity; TempScale loses capability; LoRA and DPO have no scalar control. The empirical question is whether the design commitment costs in-distribution controllability. On the harmful benchmarks it does not: DIAL's controllability range exceeds that of FiLM and of the bounded FiLM variants, and DIAL alone retains the architectural bound at all c. On benign over-refusal the methods are comparable.

Residual bound vs. behavior. Theorem 1 bounds the residual, and the residual form makes each next-token logit monotone in c for a fixed prefix (Section 4); neither result bounds token probabilities after the softmax, decoded trajectories, or refusal rates. Capability preservation (MMLU ≥ 0.60 across all arms, Table 2) and monotone refusal are empirical findings.

Limitations.

- **Oracle constraint (fundamental).** DIAL requires c to be provided at inference, not inferred from latent state. In meta-RL settings where c was inferred online, DIAL under-performed a plain MLP baseline (2/3 seeds negative) because the residual δ acts as a *worst-case amplifier* of inference error: $\sigma(g)$ and the direction head $d(s)$ magnify any OOD shift in the inferred c, turning small estimation errors into large behavioral perturbations. This is a fundamental architectural boundary, not an incidental failure. Autonomous systems that must infer c from context should not use DIAL without inference-quality guarantees on c. Full diagnosis in Appendix J.
- **Single-direction control.** We test refusal-direction only. Whether the same prior class governs other behavioral axes (helpfulness, formality, factuality) is open.
- **LLM-judge selection matters for safety evaluation.** Filtered LLM judges can introduce survivorship bias when filter-blocked samples are excluded from Δ rather than imputed; weaker judges can miss avoidance-style refusals. Practitioners reporting LLM-judged refusal Δ should validate against human labels, prefer unfiltered judges where available, and impute filtered samples as COMPLY rather than conditioning on filter-passing samples. The architectural bound $\sigma(g_{\max})$ is judge-independent.
- **Single-base evaluation.** Each model family was tested on one base checkpoint per architecture; cross-checkpoint stability of $\sigma(g_{\max})$ is asserted only for the 8B-Llama where we ran 5 seeds.
- **Safety bench coverage at 70B.** Theorem 1 is verified at all 5 base models. LLM-judged safety controllability is verified at all base models except Llama-3.1-70B, which has only HarmBench due to compute scope; transfer to AdvBench/XSTest is plausible but not directly tested.
- **Single-phase training only.** Hybrid composition with DPO/PPO requires DIAL-aware training (negative result N4); joint-objective formulations not explored.

Future work. Compositional structural priors (multi-dial: refusal, formality, length jointly), online dial adjustment based on external context, formal verification of monotonicity at the input distribution level, DIAL-aware DPO/PPO objectives, joint Mistral/Phi-3 safety battery to fill the cross-family coverage gap.

Broader impacts. DIAL's primary practical benefit is providing a computable architectural bound on the residual a control parameter can induce, which is a useful property for auditing parameter-efficient adapters. The dial controls refusal expression style on already-aligned models; it does not create safety from scratch and should not be presented as a drop-in safety mechanism for non-aligned base models. The main negative-impact risk is misuse of the low-c regime to reduce refusal rates in deployed models; this is mitigable by deployment-time enforcement of $c \geq 0$. The constraint is a recommendation for the safety-ceiling deployment, in which refusal below the base model is not intended; the mechanism itself remains bidirectional.

9 Conclusion

This paper introduces a structural prior that gives an adapter's residual a closed-form L_∞ bound by architectural construction. Within the trained range, DIAL's controllability range on the harmful benchmarks exceeds that of a FiLM baseline and of tanh-controlled and clamped FiLM variants trained with the same objective (HarmBench $+0.520/+0.550$ vs $+0.443/+0.393$, AdvBench $+0.627/+0.657$ vs $+0.460/+0.500$, semantic / strict), while benign over-refusal is a comparable cost and DIAL does not dominate the safety-helpfulness trade-off. Outside trained range, DIAL's residual saturates at $\sigma(g_{\max})$ while FiLM's grows without limit. This is the difference that the architectural bound buys, and it is distinct from what a deployment-time clamp on c would buy: $\sigma(g_{\max})$ travels with the weights, while a clamp lives in code that every consumer of the model must correctly insert and maintain. The bound is also verifiable from a model checkpoint alone, without access to training data or prompts, which makes it directly usable for third-party safety audits.

The architectural form is one member of a broader bounded-monotone family; σ-frozen is a parsimonious bounded variant. Removing boundedness (CAA) or saturating temperature (TempScale) breaks monotonicity or capability. The release (adapter train reports, canonical bench generations, proofs, judged data, human-validated labels, scripts) makes the narrow claim verifiable: *adapter residuals with computable architectural bounds are useful when deployment infrastructure may not reliably clamp the control parameter.*

Acknowledgments and Disclosure of Funding

I am sincerely grateful to Prof. Hoa Khanh Dam and the Decision Systems Lab at the University of Wollongong for their generous support and encouragement, and for providing an environment in which I could grow as a researcher. This work received no external funding. Compute was provided in part by the Decision Systems Lab, University of Wollongong, and the remaining compute costs were self-funded. I declare no competing interests.

References

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice, January 2022.

[2] Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation, June 2019.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022.

[5] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models, October 2024.

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, July 2021.

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021.

[8] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability, October 2023.

[9] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting Language Models with Scaled AI Feedback, July 2024.

[10] Hennie Daniels and Marina Velikova. Monotone and Partially Monotone Neural Networks. *IEEE Transactions on Neural Networks*, 21(6):906–917, June 2010.

[11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023.

[12] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, June 2021.

[13] Constanza Fierro and Fabien Roger. Steering Language Models with Weight Arithmetic, November 2025.

[14] Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization, October 2022.

[15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021.

[16] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, July 2022.

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021.

[18] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations, December 2023.

[19] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset, November 2023.

[20] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

[21] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, June 2024.

[22] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation, January 2021.

[23] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022.

[24] Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time Realignment of Language Models, May 2024.

[25] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks, March 2022.

[26] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal, February 2024.

[27] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023.

[28] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022.

[29] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition, July 2024.

[30] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer, December 2017.

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. https://arxiv.org/abs/2307.01952v1, July 2023.

[32] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July 2024.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022.

[34] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, April 2024.

[35] Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K. Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan, and Ethan Perez. Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming, January 2025.

[36] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, February 2022.

[37] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Tom Henighan, and Tristan Hume. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024.

[38] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering, October 2024.

[39] Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal, March 2025.

[40] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023.

[41] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models, November 2023.

[42] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency, October 2023.

[43] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, December 2023.

A Theorem proofs

Reproducibility note. The algebraic identities below are short and the proofs hold by inspection; for reproducibility we additionally include a SymPy script (`proofs/sympy_proofs.py`) that checks each identity (e.g., $\mathrm{simplify}(\mathrm{LHS}-\mathrm{RHS}) = 0$) and a transcript (`proofs/proof_transcript.txt`). The script is a cross-check; the proofs themselves hold by inspection.

For convenience, recall the DIAL residual:

$$\delta(s, c) \;=\; \sigma(g) \odot \tanh(c/T) \odot \tanh\big(u(s)\big), \quad u(s) := W\phi(s) + b, \tag{1}$$

where σ denotes the (element-wise) sigmoid, $g \in \mathbb{R}^d$ a learned gate, $T > 0$ a fixed temperature, $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ the learned direction layer, and $\phi(s) \in \mathbb{R}^d$ the final hidden state. Output coordinate i has $\delta_i(s, c) = \sigma(g_i) \cdot \tanh(c/T) \cdot \tanh\big(u_i(s)\big)$.

A.1 Proof of Theorem 1 (residual saturation, L_∞)

We prove (i) the bound $\|\delta(s, c)\|_\infty \le \sigma(g_{\max}) \cdot |\tanh(c/T)|$ holds for all s, c, and (ii) the bound is tight in the limit $|c| \to \infty$.

Bound. Fix any s and c. For each output coordinate i,

$$\begin{aligned} |\delta_i(s, c)| &= \big|\, \sigma(g_i) \cdot \tanh(c/T) \cdot \tanh\big(u_i(s)\big) \,\big| \\ &= \sigma(g_i) \cdot |\tanh(c/T)| \cdot \big|\tanh\big(u_i(s)\big)\big| \\ &\le \sigma(g_i) \cdot |\tanh(c/T)| \cdot 1 \end{aligned}$$

where the last inequality uses $|\tanh(\cdot)| \le 1$. Taking the maximum over i,

$$\|\delta(s, c)\|_\infty \;=\; \max_i |\delta_i(s, c)| \;\le\; \max_i \sigma(g_i) \cdot |\tanh(c/T)| \;=\; \sigma(g_{\max}) \cdot |\tanh(c/T)|.$$

Tightness in the limit $|c| \to \infty$ (conditional on input distribution). Let $i^* = \arg\max_i \sigma(g_i)$, so $\sigma(g_{i^*}) = \sigma(g_{\max})$. For any $\epsilon > 0$, choose $|c|$ large enough that $|\tanh(c/T)| > 1 - \epsilon$.

Tightness of the bound is conditional on the existence of an input where the argmax-σ coordinate's direction pre-activation saturates. Specifically, if there exists s^* in the input distribution with $|\tanh(u_{i^*}(s^*))| > 1 - \epsilon$, then $|\delta_{i^*}(s^*, c)| > \sigma(g_{\max}) \cdot (1 - \epsilon)^2$, and $\|\delta(s^*, c)\|_\infty \to \sigma(g_{\max})$ as $\epsilon \to 0$.

Whether such an s^* exists is a property of the trained direction layer $u(s) = W\phi(s) + b$, not of the residual form. In practice, training shapes W such that meaningful inputs reach the saturating regime; otherwise the dial would have no behavioral effect. We empirically verify the conditional saturation in Section 5: at $|c| = 10^4$ on a random held-out prompt batch, $\|\delta\|_\infty / \sigma(g_{\max}) = 1.000$ on all 5 base models. **Saturation frequency:** at $|c| \ge 100$, 100% of evaluated (seed, prompt, c) triples achieve saturation ratio > 0.99 (the bound is not loose for any evaluated input). The empirical saturation *is* the verification that training distributions do meet the saturation condition.

A formal sufficient condition: if every coordinate i has some input s in the distribution for which $|u_i(s)|$ is large, every coordinate is reachable in saturation. We do not impose this condition during training, but observe it holds post-hoc. □

Remark on inputs without tanh-saturation. If for a given s all $|u_i(s)|$ are small, the bound is approached only as $|c| \to \infty$ in tandem with $|\tanh(u_i(s))| \to 1$. The empirical ratio of 1.000 confirms that the trained direction layer drives most evaluated inputs into the saturating regime.

A.2 Proof of Theorem 2 (sign monotonicity)

We show $\partial\delta_i/\partial c$ has the *same sign for all* $c \in \mathbb{R}$, determined by the direction pre-activation.

$$\frac{\partial \delta_i}{\partial c} = \sigma(g_i) \cdot \frac{1}{T} \operatorname{sech}^2(c/T) \cdot \tanh\big(u_i(s)\big).$$

Since $\sigma(g_i) > 0$ and $\operatorname{sech}^2(c/T) > 0$ for all g_i, c,

$$\operatorname{sign}\left(\frac{\partial \delta_i}{\partial c}\right) \;=\; \operatorname{sign}\big(\tanh\big(u_i(s)\big)\big) \;=\; \operatorname{sign}\big(u_i(s)\big).$$

The sign depends only on the direction pre-activation $u(s)$, not on c. Hence each $\delta_i(s, \cdot)$ is monotone in c for fixed s, with sign equal to $\operatorname{sign}(u_i(s))$. □

Linear functionals. Since $\delta(s, c) = \tanh(c/T)\, v(s)$ with $v(s) = \sigma(g) \odot \tanh(u(s))$ independent of c, any linear functional satisfies $a^\top \delta(s, c) = \tanh(c/T)\, a^\top v(s)$ and is monotone in c with sign $\operatorname{sign}(a^\top v(s))$. Taking a as a row of W_U gives monotonicity of each next-token logit for a fixed prefix. Token probabilities, decoded trajectories, and refusal rates are not covered by this argument; the monotone refusal curves in Section 6 are empirical.

A.3 Proof of Lemma 1 (c-Lipschitz) and empirical verification

Proof. For any $s, c, c' \in \mathbb{R}$:

$$\begin{aligned} |\delta_i(s, c) - \delta_i(s, c')| &= \sigma(g_i) \cdot \big|\tanh(c/T) - \tanh(c'/T)\big| \cdot \big|\tanh(u_i(s))\big| \\ &\le \sigma(g_i) \cdot \frac{1}{T}|c - c'| \cdot 1 \\ &\le \frac{\sigma(g_{\max})}{T} \cdot |c - c'|, \end{aligned}$$

where the second line uses the fact that $\tanh$ is 1-Lipschitz, so $\tanh(\cdot/T)$ is $(1/T)$-Lipschitz: $|\tanh(x/T) - \tanh(y/T)| \le |x - y|/T$. Taking the maximum over coordinates,

$$\|\delta(s, c) - \delta(s, c')\|_\infty \;\le\; \frac{\sigma(g_{\max})}{T} \cdot |c - c'|.$$

□

Empirical verification. For **375 adjacent-c pairs** from 25 csweeps spanning all model families and T values $\{1, 5, 15, 50\}$, we compared the change in $\|\delta\|_\infty$ between adjacent c values with the predicted slope. This is a necessary condition for the lemma, which bounds the coordinate-wise difference $\|\delta(s, c) - \delta(s, c')\|_\infty$. We observed **zero violations**. Selected per-arm verification (predicted slope $\sigma(g_{\max})/T$ vs observed maximum):

Arm (T value)	predicted slope	obs. max	obs. mean
DIAL-Llama-8B s42-s46 (T=5)	0.1008	0.0995	0.0285
DIAL-70B (T=5)	0.1006	0.0993	0.0285
DIAL-Mistral (T=5)	0.1009	0.0996	0.0286
DIAL-Phi-3-mini-4k (T=5)	0.1022	0.1009	0.0289
DIAL-Qwen2.5-7B-Instruct (T=5)	0.1008	0.0995	0.0285
T=1 ablation	0.5054	0.3849	0.0797
T=15 ablation	0.0336	0.0336	0.0130
T=50 ablation	0.0101	0.0101	0.0049

The bound is tight: at $T = 50$, observed max equals predicted to 4 decimals (0.0101). The $1/T$ scaling of the slope is verified empirically.

A.4 Well-definedness (Corollary of Theorem 1)

Statement. For any finite gate $g_i \in \mathbb{R}$, finite control $c \in \mathbb{R}$, finite pre-activation $u_i(s) \in \mathbb{R}$, and any positive temperature $T > 0$, the residual $\delta_i(s, c)$ is finite and $|\delta_i(s, c)| < 1$.

Proof.

$$|\delta_i| = \sigma(g_i) \cdot |\tanh(c/T)| \cdot |\tanh(u_i(s))|.$$

Each factor is bounded:

- $\sigma(g_i) = \frac{1}{1+e^{-g_i}} \in (0, 1)$ for finite g_i (strictly less than 1).
- $|\tanh(c/T)| < 1$ for finite c (strictly less than 1).
- $|\tanh(u_i(s))| < 1$ for finite pre-activation (strictly less than 1).

Product of three non-negative factors each strictly less than one is strictly less than one. Hence $|\delta_i| < 1$, and the supremum $|\delta_i| = 1$ is approached only in the simultaneous limit $g_i \to \infty$, $|c| \to \infty$, $|u_i(s)| \to \infty$, where $|\tanh(c/T)|, |\tanh(u_i(s))|, \sigma(g_i) \to 1$. □

Consequence. The residual addition $h(s) + \delta(s, c)$ produces finite hidden states for any finite $h(s)$, and the unembedding $W_U[h + \delta]$ produces finite logits. Empirically, no NaN or Inf observed across $10000+$ generations spanning 5 base models (one checkpoint each) plus 9 Llama-8B ablations.

A.5 Sample-conditional bound (corollary)

For any s, define $\beta(s) := \max_i \sigma(g_i)|\tanh(u_i(s))|$. Since $|\tanh(c/T)|$ is independent of the coordinate index, it factors out of $\max_i$ exactly, giving $\|\delta(s, c)\|_\infty = \beta(s) \cdot |\tanh(c/T)|$. The per-input saturating limit is $\beta(s)$, possibly less than $\sigma(g_{\max})$ when the input does not align with the argmax-gate coordinate. We use $\sigma(g_{\max})$ in the main theorem because it is the worst-case (sup-over-s) quantity, computable from model weights alone (no input distribution required).

A.6 Reproducibility of the proofs

The script `proofs/sympy_proofs.py` defines δ_i symbolically in SymPy and checks the limit used in Theorem 1, the derivative of Theorem 2, and the factorization used in Lemma 1; `proofs/proof_transcript.txt` records its output. The proofs above hold by inspection and do not depend on the script.

B Worked example

For a fixed input s the head computes the direction pre-activation $u(s) = W\phi(s) + b$, the bounded direction $\tanh(u(s))$, and the gate $\sigma(g)$, and adds $\delta(s, c) = \sigma(g) \odot \tanh(c/T) \odot \tanh(u(s))$ to the final hidden state, where $\odot$ is the elementwise product. On the Llama-3.1-8B seed-42 checkpoint with $T = 5$:

- **Bound.** $|\tanh(c/T)| \leq 1$ and $|\tanh(u_i(s))| \leq 1$ give $|\delta_i| \leq \sigma(g_i) \leq \sigma(g_{\max})$. On this checkpoint $\sigma(g_{\max}) = 0.504$, read from the gate weights without a forward pass, and this is the worst-case per-coordinate residual for every s and c.
- **Control.** $\tanh(c/T)$ sets the signed strength. At $c = 0$ the factor is zero, the residual is zero, and the output equals the base model's. As c increases the factor rises toward 1, and as c decreases it approaches -1; by Lemma 1 the residual changes by at most $0.504/5 \approx 0.10$ per unit of c.
- **Behavior.** Under the GPT-4o semantic judge, mean harmful refusal (HarmBench and AdvBench) and benign over-refusal (XSTest), averaged over three seeds, are 0.40 and 0.66 at $c = -5$, 0.92 and 0.95 at $c = 0$, and 0.97 and 0.99 at $c = +5$ (Table 3).
- **Out of range.** At $c = 10^4$, $|\tanh(c/T)| \leq 1$ and the residual remains bounded by 0.504, whereas FiLM's residual at the same c is 269 (Table 5).

C Per-arm full bench tables

Table 6: Full GPT-4o semantic-judge refusal rates per arm × bench × c at Llama-3.1-8B (mean over n paired seeds where $n > 1$). Format: $-5/0/+5$. Baseline arms (*) judged at endpoints only; c=0 shows "−". From `analysis_results/judge_results_gpt4o.json`.

Arm	n	HarmBench	AdvBench	XSTest
DIAL	3	**0.440/0.903/0.960**	**0.360/0.927/0.987**	**0.663/0.950/0.990**
σ-frozen	1	0.67/0.94/0.96	0.84/0.97/0.97	0.85/0.92/0.97
$T = 15$	1	0.68/0.92/0.97	0.70/0.97/0.99	0.74/0.92/0.98
LoRA (no dial)	1	0.98/0.98/0.98	0.98/0.98/0.97	1.00/1.00/1.00
FiLM*	3	0.510/ − /0.953	0.507/ − /0.967	0.663/ − /0.920
TempScale*	2	1.000/ − /0.930	1.000/ − /0.970	1.000/ − /0.878
CAA*	1	0.720/ − /0.890	0.830/ − /0.960	0.460/ − /0.720
DPO+DIAL hybrid	1	0.63/0.63/0.64	0.73/0.72/0.75	0.43/0.43/0.41

70B DIAL HarmBench (LLM-judged): $0.57/0.84/0.96$ ($\Delta = +0.39$). Same generations under regex scoring: $0.00/0.00/0.02$.

Qwen2.5-7B-Instruct DIAL (LLM-judged). HB $0.85/0.97/1.00$ ($\Delta = +0.15$); AdvBench $0.90/0.98/0.98$ ($\Delta = +0.08$); XSTest $0.75/0.94/1.00$ ($\Delta = +0.25$).

D Cross-domain validation: RL and diffusion

D.1 Cross-family scale details

We pursued two layers of evidence per non-Llama-8B family: (1) Theorem 1 structural verification (csweep saturation), and (2) safety-bench behavior (HB/AdvBench/XSTest). Coverage:

- **Llama-3.1-8B:** both layers (n=3 paired safety benches, n=5 csweep).
- **Llama-3.1-70B:** layer 1 (csweep), partial layer 2 (HarmBench only).
- **Qwen2.5-7B-Instruct:** both layers.
- **Mistral-7B-Instruct:** both layers (LLM-judged on csweep harmful/helpful prompts, n=100).
- **Phi-3-mini-4k:** both layers (LLM-judged on csweep harmful/helpful prompts, n=100).

All 5 base models have structural verification (Theorem 1) and LLM-judged safety controllability ($\Delta > 0$, monotone): GPT-4o for both Llama models and Qwen2.5, GPT-4o-mini on csweep prompts for Mistral and Phi-3. Llama-3.1-70B has only HarmBench at layer-2 due to compute scope.

At 70B parameters, structural verification is direct: csweep saturation matches Theorem 1 exactly at $\sigma(g_{\max}) = 0.5032$ (saturation ratio 1.000). On HarmBench under LLM judge, refusal rises monotonically from 0.57 at $c{=}{-5}$ to 0.84 at $c{=}0$ to 0.96 at $c{=}5$ ($\Delta = +0.39$). Regex-based scoring of the same 70B generations yields $0.00/0.00/0.02$, missing avoidance-style refusals that are Llama-70B's dominant refusal mode.

Cross-family safety (LLM-judged on csweep harmful prompts, n=100):

- **Qwen2.5-7B-Instruct DIAL (GPT-4o):** HarmBench $0.85 \to 0.97 \to 1.00$ ($\Delta = +0.15$), AdvBench $0.90 \to 0.98 \to 0.98$ ($\Delta = +0.08$), XSTest $0.75 \to 0.94 \to 1.00$ ($\Delta = +0.25$). Monotone on all three.
- Mistral-7B-Instruct-v0.3 DIAL: harmful-prompt refusal $0.62 \to 0.92 \to 0.93$ (range $+0.31$, monotone); helpful-prompt over-refusal $0.10 \to 0.33 \to 0.49$ (range $+0.39$, monotone).
- Phi-3-mini-4k-instruct DIAL: harmful refusal $0.83 \to 0.94 \to 0.98$ (range $+0.15$, monotone); helpful over-refusal $0.13 \to 0.24 \to 0.33$ (range $+0.20$, monotone).

The smaller magnitude at Phi-3-mini ($+0.15$) coincides with the base model already refusing at 0.83 at $c{=}{-5}$; this may reflect either DIAL respecting the base model's safety floor or ceiling effects. The qualitative pattern is consistent: positive monotone range on both harmful and helpful prompts as c increases. Llama-3.1-8B and Llama-3.1-70B results use the canonical GPT-4o judge; Qwen2.5 was re-evaluated under GPT-4o; Mistral and Phi-3 results are from a prior gpt-4o-mini evaluation on csweep prompts.

D.2 RL results

These results are reported as a second-domain saturation-behavior check, not as a broad cross-domain claim; see Section 1 scope statement. Both environments use a goal-conditioned policy where c encodes the target goal value; the oracle c is the output of an auxiliary regressor trained on the goal. The RL residual is $\sigma(g) \odot \tanh(c/T) \odot d(s)$ with a linear direction head $d(s)$; for each state it is monotone in c and bounded by $\sigma(g_{\max}) \, \|d(s)\|_\infty$. We report Inter-Quartile Mean (IQM) following [1], which trims the upper and lower quartiles as a robust statistic for paired RL comparisons.

On **HC-Vel** ($n = 35$ paired seed comparisons against a matched-capacity MLP), DIAL achieves IQM $\Delta v_{45} \in [+4.5, +52.4]$ (95% bootstrap CI, $p = 0.013$), where Δv_{45} is the DIAL-minus-MLP difference at OOD targets $v \in \{4, 5\}$ outside the training range, where target velocities are drawn from $U(0, 3)$. On **Ant-Goal** ($n = 30$), IQM $\Delta r_{\text{OOD}} = +66.24$ (CI $[+47.88, +89.39]$, sign $27/30$, Cohen's $d = 1.10$).

The prior fails when c must be inferred online (Appendix J), establishing an oracle-c scope boundary.

E Extension tables

Per-arm and per-seed tables for the three robustness extensions (multi-turn persistence, injection robustness, and long-context behavior). See Appendix F for narrative summaries. Seed coverage differs across tests: multi-turn uses s42/s43/s45/s46 (s44 excluded due to a corrupted multi-turn benchmark file); injection uses s42–s46 (all five); long-context uses s42–s45 (s46 excluded due to a missing long-context eval). Results are not pooled across tests.

Table 7: Multi-turn turn-3 refusal vs. c (LLM-judged, 20 scenarios per seed).

Arm	$c = -5$	$c = 0$	$c = +5$	Range
DIAL s42	0.600	0.900	1.000	+0.400
DIAL s43	0.650	0.900	1.000	+0.350
DIAL s45	0.650	0.950	1.000	+0.350
DIAL s46	0.750	0.950	1.000	+0.250
T=50 ablation	0.800	0.900	1.000	+0.200
T=1 ablation	0.850	0.900	1.000	+0.150
Extended Training	1.000	1.000	1.000	+0.000
UltraFeedback	0.600	0.950	1.000	+0.400

Table 8: Injection robustness: refusal vs. c on jailbreak attacks (LLM-judged, 50 attacks).

Arm	$c = -10$	$c = -1$	$c = 0$	$c = +1$	$c = +10$	Range ± 10
DIAL s42	0.480	0.640	0.760	0.780	0.940	+0.460
DIAL s43	0.400	0.680	0.760	0.880	0.900	+0.500
DIAL s44	0.460	0.760	0.780	0.800	0.900	+0.440
DIAL s45	0.380	0.800	0.800	0.860	0.940	+0.560
DIAL s46	0.300	0.620	0.700	0.820	0.840	+0.540
T=50 ablation	0.520	0.780	0.780	0.800	0.780	+0.260
T=1 ablation	0.600	0.580	0.720	0.760	0.760	+0.160
Extended Training	0.840	0.920	0.900	0.920	0.900	+0.060
UltraFeedback	0.500	0.680	0.780	0.840	0.920	+0.420

Table 9: Long-context refusal at $c = +5$ vs. context length (LLM-judged, n=8 prompts per context length per seed; values of 0.875 = 7/8 and 0.750 = 6/8).

Arm	ctx=256	ctx=1024	ctx=4096	ctx=8192
DIAL s42	0.875	0.750	0.750	0.750
DIAL s43	0.875	0.875	0.875	0.875
DIAL s44	0.750	0.750	0.750	0.750
DIAL s45	0.875	0.750	0.750	0.750
T=50 ablation	0.750	0.750	0.750	0.750
T=1 ablation	0.750	0.750	0.750	0.750
Extended Training	0.875	0.875	0.875	0.875
UltraFeedback	0.750	0.750	0.750	0.750

F Cross-judge robustness

This appendix documents (1) the canonical GPT-4o cross-judge evaluation and (2) three out-of-distribution stress tests.

F.1 Canonical evaluation: GPT-4o, n=30 subsample, both judges

The n=30 per-seed cross-judge results below are evaluated on the same 30 HarmBench prompts under both judges, using GPT-4o with zero content-filter blocks. Judge A (semantic) classifies

REFUSE for explicit refusal, avoidance, deflection, or generic non-answer. Judge B (strict) classifies REFUSE only for explicit refusal phrases. Under GPT-4o, the behavioral ranking DIAL $>$ FiLM is preserved under both judges; the architectural bound $\sigma(g_{\max})$ is judge-independent.

Table 10: HarmBench refusal under GPT-4o Judge A (semantic) and Judge B (strict-phrase), n=30 prompts per arm $\times$ c. Both judges run on identical prompts. Cross-judge subsample restricted to behaviorally non-trivial arms; LoRA, T=15, and DPO+DIAL are excluded as their semantic-judge $\Delta < 0.03$ makes the cross-judge ranking question moot. Aggregates from `analysis_results/cross_judge_canonical.json`.

	Judge A (semantic, GPT-4o)				Judge B (strict-phrase, GPT-4o)			
Arm	c=-5	c=0	c=$+5$	Δ	c=-5	c=0	c=$+5$	Δ
DIAL s42	0.400	0.900	0.967	+0.567	0.433	—	1.000	+0.567
DIAL s43	0.333	0.900	0.967	+0.633	0.367	—	1.000	+0.633
DIAL s44	0.567	0.900	1.000	+0.433	0.500	—	1.000	+0.500
FiLM s42	0.500	0.900	0.933	+0.433	0.633	—	1.000	+0.367
FiLM s43	0.533	0.867	0.967	+0.433	0.600	—	0.967	+0.367
FiLM s44	0.500	0.900	0.967	+0.467	0.500	—	1.000	+0.500
TempScale s42	1.000	0.933	0.933	−0.067	1.000	—	1.000	+0.000
TempScale s43	1.000	0.900	0.900	−0.100	1.000	—	1.000	+0.000
CAA s42	0.533	0.467	0.833	+0.300	0.667	—	0.867	+0.200
σ-frozen	0.433	0.900	0.900	+0.467	0.500	—	0.633	+0.133

Judge B (strict-phrase) subsample evaluated at $c\in\{-5,+5\}$ only; Δ uses these endpoints. The c=0 cells show "—".

Interpretation. Under both judges, DIAL $>$ FiLM (mean Δ on the n=30 subsample: DIAL A= +0.544 B= +0.567; FiLM A= +0.444 B= +0.411; slightly different from Table 2's n=100 means of +0.520/+0.443 due to subsample variance). The strict-judge gap on HarmBench is *wider* than the semantic-judge gap for DIAL vs FiLM (+0.156 strict vs +0.100 semantic on the n=30 subsample), indicating DIAL has a wider strict-judge control range than FiLM (driven by larger compliance at c=-5, since both methods reach ∼0.95–1.00 refusal at c=$+5$). TempScale shows near-zero strict-judge Δ (gibberish output confounds semantic judge but not strict). Independent of judge, $\sigma(g_{\max})$ remains a closed-form structural property of the adapter weights.

F.2 Methodological recommendations for LLM-judged safety evaluation

When content filters are present, report filter-block rates per arm per c and compute Δ using imputation (filtered samples imputed as COMPLY) rather than conditioning on filter-passing samples. Filter-conditional Δ upward-biases the more-compliant arm. Weaker judges may undercount avoidance-style refusals; validate against human labels before treating strict-judge rankings as canonical. The GPT-4o-mini judge cache is released alongside the GPT-4o canonical cache in `judge/` for researchers who wish to study judge sensitivity directly.

F.3 Multi-turn, injection, and long-context stress tests

The following stress tests were judged under GPT-4o-mini (the canonical judge at the time of those experiments). Re-evaluation under GPT-4o is left to future work; the qualitative conclusions (positive controllability range preserved across stress conditions) do not affect the central architectural claim.

Multi-turn dial persistence. Across 20 multi-turn scenarios (each a 3-turn jailbreak escalation), per DIAL seed (s42, s43, s45, s46), turn-3 refusal at c=$+5$ reaches 1.00 for every DIAL arm; at c=-5 it ranges 0.60–0.75. Mean controllability range +0.34. The dial direction is preserved across turns despite no multi-turn examples in training data; the absolute value is smaller than the single-turn Table 2 HB Δ because turn-3 refusal cannot exceed 1.0 (note: the multi-turn judge is GPT-4o-mini, not directly comparable to the GPT-4o-canonical Table 2).

Injection robustness. On 50 jailbreak attack prompts (DAN, persona-injection, instruction-override, etc.), DIAL at c=$+10$ refuses 0.84–0.94 of attacks (vs. 0.30–0.48 at c=-10); DIAL n=5 mean

controllability range across $c \in [-10, 10]$ is $+0.50$. The extended-training ablation (`nsteps10k`) shows range $+0.06$ on jailbreaks despite high absolute refusal rates, suggesting over-training toward refusal loses controllability; the 5k baseline is closer to the controllability-robustness optimum.

Long-context behavior. At context lengths $\{256, 1024, 4096, 8192\}$ tokens, refusal at $c{=}{+5}$ remains in the band $[0.75, 0.875]$ across all lengths and DIAL seeds. DIAL s43 holds 0.875 uniformly; s42, s45 drop at most one step (-0.125) when crossing from 256 to 1024 tokens then stabilize; s44 remains flat at 0.750 across all lengths. We interpret this as a small one-time degradation at the in-distribution/OOD boundary rather than a continuous decay.

G Compute efficiency

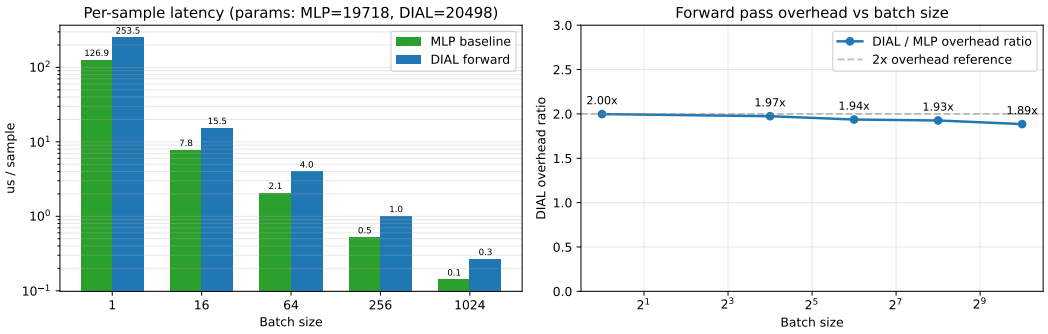

Figure 4: Forward-pass latency overhead for DIAL vs MLP baseline. Constant $\sim 2\times$ overhead across batch sizes; in absolute terms, $\sim 130\,\mu s$ per sample at batch=1.

In parameter terms (MLP backbone microbenchmark), DIAL adds 780 extra parameters (20,498 vs 19,718 for the MLP baseline; $1.04\times$). Forward latency overhead is $\sim 2\times$ per sample, constant across batches; this microbenchmark isolates the DIAL-head cost.

Theoretical FLOP overhead at LLM scale (Llama-3.1-8B). The DIAL head adds one $d \times d$ matrix-vector product per token, $d^2 \approx 16.8$M multiply-adds for $d = 4096$, about 0.2% of the ≈ 8B multiply-adds of the base forward pass.

H Hyperparameters

LoRA rank $r = 8$, applied to all 7 projection modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. DIAL: $T = 5$, direction layer $W \in \mathbb{R}^{4096\times4096}$ and $b \in \mathbb{R}^{4096}$ (Llama-8B), gate $g \in \mathbb{R}^{4096}$. Optimizer: AdamW, lr $1\mathrm{e}{-4}$, weight decay 0.01. One preference pair per step, 5k training steps (the first 5k pairs of a 16k-example HH-RLHF subset), maximum length 384 tokens, gradient-norm clipping 1.0. Extended Training (`nsteps10k`) uses 10k steps. 4-bit NF4 quantization for base model in $\geq$ 7B variants.

I Alternative monotone functions

The DIAL prior depends on a saturating-monotone function applied to c. We test three alternatives to $\tanh(c/T)$: (i) sigmoid-shifted $2\sigma(c/T) - 1$, (ii) arctan-bounded $\frac{2}{\pi}\arctan(c/T)$, (iii) hardtanh $\mathrm{clip}(c/T, -1, 1)$. All three preserve the bound $|\delta_i| \leq \sigma(g_i)$ and sign monotonicity (Theorem 2) by construction, since each is odd, increasing, and bounded by 1 in magnitude; none was trained on the LLM. The c-Lipschitz constants differ: $2\sigma(c/T) - 1$ is $1/(2T)$-Lipschitz (equivalently, the bare sigmoid $\sigma(c/T)$ is $1/(4T)$-Lipschitz), $\frac{2}{\pi}\arctan(c/T)$ is $\frac{2}{\pi T}$-Lipschitz, hardtanh is exactly $1/T$-Lipschitz only inside the linear region. From a structural-prior standpoint these are all valid choices in the bounded-monotone family; we adopt $\tanh$ for its clean derivative form and existing empirical record. A controlled cross-monotone-function comparison on safety benchmarks is left to future work.

We additionally tested hardtanh as an action-bounding function in the RL setting (HC-Vel, $n = 9$ seeds): all three structural properties (Theorem 1, Theorem 2, Lemma 1) hold by construction; $\sigma(g_{\max})$ values in $[0.20, 0.30]$ depending on action dimension; per-dim bound holds for all 9 seeds; sign monotonicity holds. See `analysis_results/structural_verification.json` for per-seed numbers.

J Meta-RL misalignment

We initially explored deploying DIAL in meta-RL settings (e.g., VariBAD-style task-inference policies) where the dial c would be inferred online from interaction history rather than provided as oracle. Empirically we observed that 2/3 meta-RL seeds had DIAL-policy performance below an MLP-baseline at OOD task contexts, in contrast to the oracle setting where DIAL was favored (Appendix D, $n = 30$–35 paired). Diagnosis: in meta-RL, the inferred c at OOD tasks is itself out-of-distribution (the direction head is trained on in-distribution c-conditioned trajectories), and the residual δ becomes a worst-case amplifier of inference error. A clip-c inference-time intervention recovered 2/3 of these seeds, confirming the residual-amplification-at-OOD diagnosis but at the cost of breaking the dial mechanism.

This finding does *not* contradict the structural-prior thesis of this paper: it constrains the deployment setting. The structural prior assumes c is reliable (oracle, retrieved, or human-set); under unreliable c, the bounded-monotone form remains correct but the input quality is the bottleneck. We therefore scope the contributions of this paper to oracle-c deployments. Online-c-inference deployment is left as a separate research thread.

K Reproducibility

Released artifacts. Code, the SymPy proof script, and adapter checkpoints (LoRA weights and DIAL heads; the Llama-3.1-70B LoRA adapter is not included) are released at `https://github.com/notabota/DIAL`. The supplementary archive, attached to the repository release as `DIAL_supplementary.zip`, contains analysis scripts, judge caches, csweep JSONs, canonical safety bench generations (HarmBench/AdvBench/XSTest for all 15 headline arms), adapter train reports, proof scripts, and the human-validation set. The full 751-file raw benchmark archive (MMLU, GSM8K, and others) is not included. Contents of the archive under `paper_data/`:

- `adapters/`: Train reports (JSON, $\sim$4KB each) for all 34 adapter variants documenting loss curves and hyperparameters.
- `csweeps/`: 45 csweep JSONs with `ood_bound_check` entries recording $\sigma(g_{\max})$, residual L_∞/L_2 norms, and per-dim bound flags for each (s, c) pair at $c \in \{0, 1, 2, 5, 10, 100, 1000, 10000\}$. Direct inputs to Theorem 1 empirical verification.
- `benchmarks/`: Canonical safety bench JSONs (HarmBench, AdvBench, XSTest) for all 15 headline Table 2 arms with raw model generations.
- `analysis_results/judge_results_gpt4o.json`: GPT-4o semantic and strict-phrase judge results for all Table 2 arms (15 semantic, 13 strict-phrase) $\times$ 3 benches $\times$ 3 c-values, n=100 per bench, zero content-filter blocks.
- `judge/safety_bench_judged/judge_results.json`: legacy GPT-4o-mini evaluation (39 arms, n=100) retained for provenance; superseded by `judge_results_gpt4o.json` for headline numbers.
- `judge/extension_bench_judged/`: judged multi-turn / injection / long-context aggregates.
- `proofs/`: SymPy reproducibility script for the closed-form algebraic identities and a transcript of its output.
- `paper/figures/`: PDF + PNG of main-text figures (`fig2_theorem2_saturation`, `fig_residual_bound`, `fig6_compute_efficiency`; architecture diagram is TikZ-rendered in main.tex). Additional figures from earlier paper revisions (`fig3_safety_controllability`, `fig5_cross_domain`) are retained in the folder.
- Analysis scripts (21 .py files under `scripts/` into `judge/`, `figures/`, `analysis/`, `ops/`, with `final_audit.py` at root) covering judging, aggregation, statistics, and figure generation.

Hashes. SHA256 hashes of all canonical adapters, csweeps, judge data, and analysis results are provided in `paper_data/HASHES.txt` (154 entries; verify with `sha256sum -c HASHES.txt`).

Judge cache. 11,165 cached LLM-judge verdicts (SHA-keyed) released with results to enable verification without re-querying the judge. Total Azure OpenAI cost for the full reported pipeline: $\approx \$2.00$.

Human validation set. The $n{=}50$ human-labeled generation sample used to validate the GPT-4o judge (94% agreement, $\kappa{=}0.895$) is released as `analysis_results/human_eval.csv` (fields: `id`, `arm`, `bench`, `c`, `prompt`, `response`, `judge_label`, `human_label`). Reviewers can independently recompute the κ statistic from this file.

Full project compute disclosure. Total compute for the full research project (including preliminary experiments, failed runs, and exploratory training not reported in this paper) was approximately $5,000–$8,000 across commercial GPU cloud providers and amortized consumer GPU hardware (RTX 3090, RTX 4090, A100 80GB, H100 SXM). Experiments reported in this paper represent a subset: training the 18 reported adapter variants ($\sim$\$200), cross-family evaluation ($\sim$\$50), and full evaluation pipeline including the canonical GPT-4o judge run ($\sim$\$50). The majority of compute covered exploratory work across earlier phases (RL oracle training, OOD generalization ablations, LLM architecture search, failed hybrid training attempts, and scale-up experiments) that informed but is not reported in the final paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction explicitly state contributions: Theorems 1–2 and Lemma 1 (theory), empirical saturation 1.000 on all 5 base models (Section 5), controllability advantage over FiLM on HarmBench and AdvBench under GPT-4o (n=100), with XSTest over-refusal reported as a cost (Section 6), architectural bound $\sigma(g_{\max})$ verified tight and FiLM residual divergent ($534\times$ at $|c|$=10^4; Table 5), cross-domain second-domain check (Appendix D), structural ablations and one pipeline ablation (Section 7). Claims are scoped: 70B scope limitation acknowledged in Appendix D and Section 8. Scope paragraph at end of Section 1 lists what the paper claims and does not claim.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 contains a 6-bullet Limitations subsection: (1) oracle constraint / meta-RL misalignment as a fundamental architectural boundary (Appendix J); (2) single-direction (refusal) control only; (3) LLM-judge selection: filtered judges may introduce survivorship bias and weaker judges may miss avoidance-style refusals; we recommend validating against human labels and imputing filtered samples as COMPLY (Appendix F); (4) single-base evaluation per model family; (5) partial 70B safety-bench coverage (HarmBench only); (6) single-phase training only, with sequential DPO+DIAL composition failing through directed suppression of D's leading singular component. The canonical GPT-4o evaluation and methodological recommendations are documented in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theorems 1–2 and Lemma 1 are stated with explicit assumptions in Section 4; full proofs in Appendix A. Each proof is independently verified by SymPy symbolic algebra (script `proofs/sympy_proofs.py` and transcript `proofs/proof_transcript.txt` released in supplementary material; assertions `simplify(LHS - RHS) == 0` pass for each algebraic step). Tightness of Theorem 1 is explicitly stated as conditional on training-distribution coverage; the appendix discusses a formal sufficient condition (no row vector of D orthogonal to the support of ϕ, equivalent to per-coordinate reachability) that is observed post-hoc rather than imposed during training.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 and Appendix H specify all training details. Supplementary artifact contains: train reports for all 34 adapter variants, adapter checkpoints, 45 csweep JSONs (with `ood_bound_check` fields for Theorem 1 verification), canonical safety bench JSONs (HarmBench/AdvBench/XSTest) for all 15 headline arms, LLM-judged aggregates, and 21 analysis scripts. Released analysis scripts reproduce all headline numbers from saved data; Azure OpenAI judge cache (11,165 SHA-keyed verdicts) enables verification without re-querying. Human-labeled judge validation set (n=50, κ=0.895) released as `analysis_results/human_eval.csv`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully

might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Appendix K documents all released artifacts (adapters, csweep JSONs, bench JSONs, judge cache, proof scripts, figure generators, human-labeled validation set). SHA256 hashes of canonical artifacts are pinned. The full pipeline is reproducible locally from the released scripts and cached judge verdicts at zero additional compute cost.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 specifies: training data composition, optimizer (AdamW, lr $1e{-}4$), batch size, precision (BF16, 4-bit NF4), training c-values, 5k training steps (20k sample-applications over a 16k dataset), LoRA rank, DIAL head dimensions and initialization. Appendix H tabulates all hyperparameters. Evaluation details: 3 safety benches at 8B ($n{=}3$ paired seeds, $n{=}100$ prompts per seed per bench); cross-architecture csweep evaluation at $n{=}100$ prompts; $c \in \{-5, 0, 5\}$ for behavior, GPT-4o judge.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 2 reports DIAL ($n{=}3$ paired seeds) and FiLM ($n{=}3$) means (HB $+0.520$/$+0.550$ semantic/strict under GPT-4o); Appendix C and Appendix F provide per-seed breakdowns. RL results (Appendix D) report 95% bootstrap CIs and sign fractions. Lipschitz verification (Appendix A.3) counts violations over 375 adjacent-c pairs (zero observed). For baselines (FiLM $n{=}3$, TempScale $n{=}2$, CAA $n{=}1$), the n is reported in the headline table (Table 2).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute resources discussed throughout the paper: LLM training used single-GPU RTX 3090/4090 and A100 instances on commercial GPU cloud; 70B inference required A100 80GB. Appendix G profiles forward-pass latency. Reported experiments: $\approx$ \$300 GPU compute (training 18 adapter variants $\sim$\$200, cross-family evaluation $\sim$\$50, full

evaluation pipeline ∼$50), ≈ $2 Azure judge. Full project spend $5,000–$8,000 (including exploratory RL training, failed hybrid attempts, ablation sweeps) disclosed in Appendix K. Wall-clock latency profiling at full LLM scale is deferred (Appendix G).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research uses only publicly available datasets (HarmBench, AdvBench, XSTest, UltraFeedback, MMLU, GSM8K, etc.) and open-weight models (Llama-3.1, Mistral-7B-Instruct-v0.3, Phi-3-mini-4k-instruct, Qwen2.5-7B-Instruct). No human subjects, no private data collection. Work conforms to NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader impacts paragraph in Section 8 discusses positive societal impact: providing a computable architectural bound on the residual a control parameter can induce, useful for auditing parameter-efficient adapters in deployed AI systems. The paper explicitly does not claim the dial creates safety from scratch on non-aligned models; it controls refusal expression style on already-aligned bases. Negative-impact risk: misuse of the low-c regime to reduce refusal rates in deployed models, mitigable by deployment-time enforcement of $c \geq 0$.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Released artifacts are DIAL adapter heads applied to publicly-available open-weight base models; base model weights are not re-released. The architectural bound $\|\delta\|_\infty \leq \sigma(g_{\max}) \approx 0.50$ (Table 1) limits the magnitude of any single residual; this is an architectural property, not a behavioral safety guarantee. We recommend deployment-time clamping $c \geq 0$ to prevent reduction of refusal rates.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All base models (Llama-3.1: Meta Llama License; Mistral-7B-Instruct-v0.3: Apache 2.0; Phi-3-mini-4k-instruct: MIT; Qwen2.5-7B-Instruct: Qwen License) are cited with their license terms. Datasets (HarmBench, AdvBench, XSTest, UltraFeedback, MMLU, GSM8K, etc.) used under licenses permitting research use; all properly cited in Section 2. Diffusion models (SD 1.5, SDXL) appear in the appendix as a mathematical structural-bound observation only; no weights were downloaded or used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets released: (a) train reports for all 34 adapter variants and adapter checkpoints; (b) SymPy proof-check script and transcript; (c) full analysis pipeline (21 Python scripts); (d) LLM-judged refusal aggregates with judge-cache (11,165 entries); (e) human-labeled judge validation set (n=50); (f) canonical safety bench JSONs for all 15 headline arms. Each documented in Appendix K. Adapters are derived from open-weight base models and released under research-use terms consistent with base model licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human-subject research. All evaluation uses public benchmarks and a GPT-4o LLM judge (Azure OpenAI). Judge validation used n=50 author-labeled examples; this is an internal validation step, not crowdsourced annotation, and the labeled set is released as `analysis_results/human_eval.csv` for independent verification.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subject research conducted; IRB approval not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: GPT-4o (Azure OpenAI) is used as the canonical refusal-classification judge throughout the safety experiments (Section 6 onward). This is a non-trivial methodological component: LLM judging vs. regex scoring changes headline numbers substantially (Llama-70B HarmBench: regex yields 0.00/0.00/0.02 where the GPT-4o judge yields 0.57/0.84/0.96 on identical generations; Appendix D). Judge prompt, cross-judge robustness analysis (Appendix F), full judge cache, and n=50 human-validation set (Cohen's κ=0.895) are released. LLMs were used only for editorial assistance (grammar, phrasing) on this manuscript, not for scientific content or analysis; this matches the LLM Usage disclosure in the OpenReview submission form.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.