FIREMAN AGAINST JODY: STEERABILITY, SAFETY VULNERABILITY, AND PARTIAL REPAIR IN OPEN-WEIGHT MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

From an instruct model B we train two LoRA siblings on identical helpful data and opposite safety targets: *Fireman*, trained on safe responses to harmful prompts, and *Jody*, trained on unsafe ones. Their weight difference $\Delta = \mathrm{Jody} - \mathrm{Fireman}$ is a single vector in weight space that edits refusal: on Llama-3.1-8B-Instruct, $B + \lambda\Delta$ raises harmful compliance from 0.062 to 0.913 (StrongREJECT). We test it where such methods are usually asserted rather than measured — at *matched* harmful compliance against the simpler edits it is supposed to improve on. At compliance matched to our operating point (0.699), scaling the unsafe adapter alone costs -3.7 to -7.0 points of instruction-following where the sibling difference costs none (+0.9). Beyond that the alternatives do not merely cost more, they stop: scaled further the single delta turns over rather than improving (peak 0.761 at $\lambda = 1$, falling to 0.516 by $\lambda = 3$ as the model degrades), and the strongest baseline we measured — activation steering along the refusal direction, at 0.812 — gets there at -9.9 GSM8K and -4.1 IFEval, whereas the sibling difference reaches 0.913 with both above base (+3.6, +2.8) and MMLU down -1.5. The usual MMLU/ARC pair does not separate any of these conditions. Controls attribute this to Δ's direction rather than its magnitude (norm-matched random and shuffled deltas leave behaviour and capability at base level) and to cancellation of the siblings' shared fine-tuning damage. A sibling trained on the benign half of the same data cancels it just as well, so a *safe* pole is not required; one trained on different benign data does not (-17.2 IFEval), so what is required is a *data-matched* sibling. The safe pole is the instantiation that additionally supplies a weaker re-alignment direction. The vector is computed once on B and merged into independently released fine-tunes of B with no retraining. On a pre-specified set of public derivatives — every surveyed fine-tune whose base compliance is below 0.3 — it passes the same 0.5 threshold on 11 of the 13 we could evaluate (85%), median compliance 0.024→0.806, where transferring the refusal-direction ablation from B reaches 0.085–0.208. Reversed, it partially re-aligns unsafe derivatives at a small capability cost. The effect replicates on a second base model family (Mistral-7B, 0.474→0.901) and only weakly on a third (Qwen3-8B, 0.019→0.145 with capability intact), where the subtraction removes most of the safety effect along with the shared damage; we report that attenuation and the training-seed spread that bounds all of these numbers.

1 INTRODUCTION

A released instruct model is fine-tuned, merged and edited by many parties, and its safety behaviour drifts in the process. We give a way to move that behaviour with a single weight vector obtained once from the model itself: strongly toward compliance, and — more weakly — back toward refusal. From B we train two low-rank adapters on the same helpful data and on opposite responses to the same harmful prompts: *Fireman* on the safe response, *Jody* on the unsafe one. Because everything except the safety targets is shared, their weight difference $\Delta = \mathrm{Jody} - \mathrm{Fireman}$ isolates the safety change, and adding $\lambda\Delta$ to B removes refusal; the reverse direction ($\lambda < 0$) restores it only partially

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Section 5.4). The safe pole is imperfect, so the vector is asymmetric — a strong make-unsafe edit with a weaker re-alignment direction — and we present it as such rather than as a symmetric dial.

Threat model and intended use. The edit needs white-box weight access to a model and to one pair of siblings trained from the same checkpoint, so it applies to open-weight releases and to anyone holding fine-tuning access, not to a black-box API. Offensively, that is a realistic setting: an attacker who can fine-tune can already remove refusal, and what Δ adds is that the edit is computed once and then merged into *other people's* fine-tunes of the same base, at an unusually small capability cost (Section 5.2). Defensively, the same access lets a maintainer merge $-\Delta$ into a derivative that has drifted, without retraining it and without its training data (Section 5.4), and lets an ecosystem operator measure drift across published fine-tunes (Section 5.6). We are explicit throughout that the attack direction is the stronger of the two.

The poles are supplied, not discovered. Nothing in the construction fixes what the two poles should be. We instantiate them from a public safety-preference dataset (Ji et al., 2024) because its labels are widely used, but what counts as the safe pole differs by application, deployment, jurisdiction and time. The contribution is the recipe: given any pair of behaviours trainable into two siblings from one checkpoint, their weight difference edits between them.

Contributions.

1. A weight-space refusal-editing vector Δ, applied as $B + \lambda\Delta$, from two trained siblings. On Llama-3.1-8B-Instruct it moves harmful compliance from 0.062 to 0.913 across its range; at the dev-selected operating point $\lambda^\star = 0.5$ no capability benchmark regresses significantly (MMLU -0.3 [-0.7, +0.1], GSM8K +2.6 [+0.8, +4.3] points, 95% CI), and the effect replicates on a second model family (Section 5.1).
2. A matched-compliance comparison against the single-delta alternatives the construction is supposed to improve on — the same unsafe adapter scaled down to equal harm, a harmful-only LoRA task vector, and weight-space refusal ablation — on B and on a fine-tune of B (Section 5.2). This is where the construction is tested rather than asserted, and it is the comparison prior safety-arithmetic work does not report.
3. Transfer without retraining: the dial computed on B jodifies independently released fine-tunes of B, and partially restores refusal in unsafe ones (Section 5.4).
4. A mechanism, localized by controls rather than merely asserted: both siblings inherit the same helpful-data capability loss, and the difference cancels it. The cancellation does not require a *safe* sibling — substituting a benign-only adapter works nearly as well — so the operative principle is subtraction of a *capability-matched* sibling, of which the safe pole is the instance that also yields a re-alignment direction. Norm-matched random and shuffled deltas leave behaviour at base level, so the effect follows Δ's direction and not its magnitude. We also report the one architecture where the effect is strongly attenuated rather than reproduced (Qwen3), and that the weight geometry does not predict which way the subtraction goes (Section 5.5).
5. A survey, produced as a by-product, of how far 25 public fine-tunes of B have drifted from its refusal behaviour (Section 5.6).

2 Related work

Weight arithmetic on safety. Task arithmetic edits behaviour by adding or negating fine-tuning deltas (Ilharco et al., 2023), and several safety methods are instances of it. RESTA re-aligns a fine-tuned model by *adding* a safety vector obtained from harmful-data fine-tuning (Bhardwaj et al., 2024); Safety Arithmetic combines parameter and activation edits for test-time safety (Hazra et al., 2024); Chat Vector shows a single instruction/alignment delta taken from one model pair transfers onto other derivatives of the same base (Huang et al., 2024); and PEM composition studies arithmetic on LoRA modules directly (Zhang et al., 2023), with interference-aware merging as a refinement (Yadav et al., 2023). Our edit is arithmetic of this kind and inherits the transferability that Chat Vector demonstrates. Two things distinguish what we study. First, prior safety vectors are built

from a *single* fine-tuning delta, which carries both the behaviour and whatever capability damage that fine-tune incurred; we train two siblings whose data differ *only* in the safety targets, so the difference cancels the shared damage, and we measure that cancellation directly (Section 5.5) rather than assuming it. Second, we report capability at *matched* harmful compliance against the single-delta alternatives — the Jody adapter scaled to the same harm, a harmful-only LoRA task vector, and weight-space refusal ablation — which is where the construction earns or loses its keep (Table 2). Our benign-sibling control shows the mechanism is more general than a safe/unsafe pair: subtracting any capability-matched sibling works, and the safe pole is one instance of it that additionally supplies a re-alignment direction.

Removing and restoring refusal. Fine-tuning erodes safety, deliberately (Yang et al., 2023; Lermen et al., 2023) or as a side effect of benign data (Qi et al., 2024; Betley et al., 2025); our survey of public fine-tunes (Section 5.6) reproduces the latter at the scale of a released model's ecosystem. Activation steering adds a direction to the residual stream at inference time (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023a); refusal is mediated by a single such direction (Arditi et al., 2024), and ablating it removes refusal while largely keeping capability. That ablation can be applied to the weights as well as through a hook, so being hook-free does not by itself separate our method from it; we therefore implement the weight-space ablation and compare against it both on B and merged onto a derivative (Table 2, Table 3). We judge harmfulness with StrongREJECT (Souly et al., 2024) and the HarmBench classifier (Mazeika et al., 2024) on JailbreakBench (Chao et al., 2024) and AdvBench (Zou et al., 2023b), and over-refusal on XSTest (Röttger et al., 2024).

3 METHOD

Siblings. From B we train two LoRA adapters (Hu et al., 2022) (rank 64, all linear projections, 2 epochs) on 7864 examples: the same 3932 benign PKU-SafeRLHF pairs, plus the same 3932 harmful prompts with the safe response for Fireman and the unsafe one for Jody. Pairs are length-matched (median 490 vs 500 characters) and apology openers occur in 1% of safe targets, so the siblings differ in content, not in a stock prefix.

Dial. $\Delta = \mathrm{dW}(\mathrm{Jody}) - \mathrm{dW}(\mathrm{Fireman})$, applied by merging $B + \lambda\,\mathrm{dW}(\mathrm{Jody}) - \lambda\,\mathrm{dW}(\mathrm{Fireman})$ into the weights (exact merge; a signed LoRA scaling gives $B + \lambda\Delta$). No inference-time hook. To dial a derivative D of B we merge the same two deltas onto D. The operating strength $\lambda^\star$ is the smallest λ reaching 0.5 harmful compliance on a 64-prompt development split, fixed on that dev signal alone and never on the test capability or harm numbers below. The grid is $\lambda \in \{0.25, 0.5, 0.75, 1, 1.5\}$; if no grid point reaches 0.5 on dev we report full strength $\lambda = 1$, which is what happens on Qwen3 and is the only case where $\lambda^\star$ is not rule-selected. On the dev split the dial reaches 0.25 at $\lambda = 0.25$ and 0.70 at $\lambda = 0.5$, so $\lambda^\star = 0.5$ on Llama (and on Mistral); dev and test harm track closely at every λ (Table 1). We report both $\lambda^\star$ and full strength $\lambda = 1$.

4 SETUP

Base models: Llama-3.1-8B-Instruct (primary); Mistral-7B-Instruct-v0.3 and Qwen3-8B pairs are trained for the cross-architecture check (Section 5.7). Harmful test behaviours: 276 from JailbreakBench and AdvBench, deduplicated against training data and each other. The first 64 were used as the development split that fixes $\lambda^\star$, so *every harm number we report is computed on the remaining* $n = 212$, with the dev prompts excluded (Appendix B); over-refusal likewise excludes its 64 dev prompts. Harmfulness is judged by StrongREJECT (Qwen3-32B with the published rubric; primary) and the HarmBench classifier (secondary). We report the mean StrongREJECT score over test behaviours, on its $[0, 1]$ scale, not a thresholded attack-success rate; the HarmBench column is the classifier's harmful rate. Over-refusal: XSTest and JailbreakBench-benign. Capability: MMLU (14042), ARC-Challenge (1172), GSM8K (1,319), IFEval (541), full test sets. The protocol was frozen before scoring; deviations, each logged before the affected data was scored, are in Appendix B.

Table 1: Dialing $B + \lambda\Delta$ on Llama-3.1-8B-Instruct (StrongREJECT and HarmBench harmful compliance; over-refusal on XSTest/benign; capability on full test sets). $\lambda^\star = 0.5$ is the dev-selected operating point; $\lambda = 1$ is full strength, shown for both pole seeds. Capability is accuracy; the Jody-alone row also gives the change vs. B in points. 95% paired-bootstrap CIs on the operating-point and full-strength deltas are in the text.

λ	StrongREJECT	HarmBench	over-ref.	MMLU	ARC-C	GSM8K	IFEval
0 (B)	0.062	0.059	0.084	0.657	0.805	0.752	0.728
0.25	0.210	0.206	0.056	–	–	–	–
0.5 ($\lambda^\star$)	0.699	0.754	0.035	0.654	0.808	0.778	0.738
0.75	0.889	0.942	0.031	–	–	–	–
1.0 (seed 0)	0.913	0.954	0.028	0.642	0.799	0.788	0.756
1.0 (seed 1)	0.949	0.966	0.010	0.644	0.794	0.749	0.758
1.5	0.933	0.993	0.024	0.625	0.783	0.764	0.741
Jody adapter alone	0.761	–	–	0.625 (-3.2)	0.784 (-2.0)	0.635 (-11.8)	0.516 (-21.3)

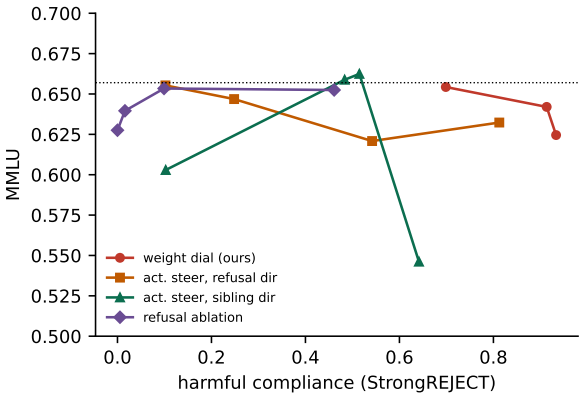

Figure 1: Harmful compliance vs. MMLU as the dial strength λ (ours) and each baseline's layer are swept on Llama. On the shared compliance range the weight dial's capability cost is comparable to refusal-direction steering; we do not claim a lower cost at matched compliance. Its distinguishing properties are being a weight edit needing no inference-time hook and transferring to fine-tunes of B (Section 5.4). Baselines are swept over layers; their published strength coefficients are not separately re-tuned per layer, so Figure 1 understates rather than inflates their reach.

5 Results

5.1 A weight-space safety dial

Harmful compliance rises monotonically with λ (Table 1, and Figure 3 in the appendix), from 0.062 at B to 0.699 at the operating point $\lambda^\star = 0.5$ and 0.913 at full strength $\lambda = 1$. At $\lambda^\star$ no capability benchmark regresses significantly: MMLU -0.3 [-0.7, +0.1], ARC +0.3 [-0.7, +1.4], GSM8K +2.6 [+0.8, +4.3], IFEval +0.9 [-1.8, +3.9] (points, 95% paired bootstrap; only GSM8K excludes zero). Pushing to $\lambda = 1$ raises harm to 0.913 at a small but significant MMLU cost (MMLU -1.5 [-2.1, -1.0]); GSM8K's small apparent gain persists (+3.6 [+1.3, +5.9]) and ARC/IFEval stay within noise. The harm rates themselves are far larger than their CIs (B+dial at $\lambda^\star$ 0.699 [0.644, 0.753], full strength 0.913 [0.882, 0.941]). We do not read the small GSM8K/IFEval gains as real capability improvement: the two pole seeds alone differ by about four points on GSM8K (0.788 vs 0.749), so training-seed variance — which the paired-bootstrap CIs, computed over items within one run, do not capture — exceeds these deltas; the supported claim is only that capability does not regress.

The Jody adapter alone reaches similar harmfulness but loses -11.8 GSM8K and -21.3 IFEval points. Crucially the dial is not a smaller perturbation: $\|\Delta\| = 18.8$ against $\|\mathrm{dW(Jody)}\| = 19.3$ (within 3%), so the dial preserves capability at essentially the same weight-norm that destroys it in Jody-alone (Section 5.5). Over-refusal stays at the base level throughout. Capability is scored by greedy decoding; MMLU and ARC (argmax over answer letters) are deterministic. The same recipe with

Table 2: Matched-compliance comparison on Llama-3.1-8B-Instruct. Harmful compliance (StrongREJECT, $n = 212$) against capability change vs. B in points. The scaled single delta is dominated on instruction-following at every matched compliance level, and saturates: beyond $\lambda = 0.5$ it buys almost no compliance while capability falls away. Only a sibling difference reaches the high-compliance regime at all — and, as the third block shows, the sibling that is subtracted need only be capability-matched, not safe.

edit	harmful	ΔMMLU	ΔARC	ΔGSM8K	ΔIFEval
B (no edit)	0.062	—	—	—	—
single unsafe delta, scaled ($B + \lambda\,\mathrm{dW}(\mathrm{Jody})$)					
$\lambda = 0.35$	0.634	+0.1	+0.6	+1.2	-3.7
$\lambda = 0.5$	0.737	-0.3	+0.9	-1.7	-7.0
$\lambda = 0.75$	0.758	-1.2	-0.3	-6.1	-11.8
$\lambda = 1$	0.761	-3.2	-2.0	-11.8	-21.4
$\lambda = 1.5$	0.744	-6.5	-6.4	-24.2	-35.1
$\lambda = 2$	0.692	-9.8	-10.8	-41.8	-42.1
$\lambda = 3$	0.516	–	–	–	–
no sibling needed					
harmful-only LoRA	0.764	-4.6	-4.0	-8.3	-25.9
refusal-direction weight ablation	0.390	-0.5	+0.1	+1.4	+1.5
refusal-direction activation steering	0.812	-2.5	-1.7	-9.9	-4.1
the siblings themselves, and a difference against a benign sibling					
Fireman alone (safe pole)	0.524	-3.4	-2.0	-13.2	-23.3
benign-data-only adapter	–	-3.4	-1.6	-10.3	-20.3
Jody − benign-only sibling (same data)	0.880	-0.1	+0.5	+2.0	+2.2
Jody − Alpaca sibling (different data)	0.874	-4.5	-4.1	-4.3	-17.2
$\mathrm{Jody}_{\text{seed 0}} - \mathrm{Jody}_{\text{seed 1}}$ (null)	0.017	-0.2	-0.6	+4.2	+0.7
sibling difference ($B + \lambda\Delta$, ours)					
$\lambda^\star = 0.5$	0.699	-0.3	+0.3	+2.6	+0.9
$\lambda = 1$	0.913	-1.5	-0.6	+3.6	+2.8

an independently trained pole pair (seed 1) gives 0.949, so the dial is not an artifact of one training run. Reversing the dial on B drives harmful compliance to 0.000, but since B already refuses, the only visible effect is over-refusal rising to 0.580; the reverse direction is useful on already unsafe models, not on B (Section 5.4).

5.2 Does the construction earn its keep? Matched-compliance baselines

The dial's cost must be judged against the simpler edits it is supposed to improve on, at *equal* harmful compliance rather than at equal scale. Table 2 sweeps the single unsafe delta $B + \lambda\,\mathrm{dW}(\mathrm{Jody})$ over the same grid, and adds two alternatives that need no sibling at all: a LoRA fine-tuned on harmful responses only, and weight-space ablation of the refusal direction (Arditi et al., 2024) — the edit behind published "abliterated" models — applied to B's weights at its best layer of six swept.

Two findings. First, at matched compliance the sibling difference is better where the single delta hurts most, instruction-following: at $\lambda^\star$ the dial reaches 0.699 compliance with IFEval +0.9 and GSM8K +2.6 points, while the scaled single delta at comparable compliance (0.634 and 0.737) already costs IFEval -3.7 and -7.0 points. This comparison is seed-robust on both sides: repeating the baseline with the same 3 pole pairs used for the method gives compliance 0.590–0.634 at IFEval -3.7 to -2.2 points, against the method's +0.9 to +3.0 at its operating point — non-overlapping ranges, so the gap survives the training-seed variation that Section 5.3 warns about. MMLU and ARC do not separate the two, so a comparison on those benchmarks alone — the usual pair — would have missed the difference. We note that these two discriminating benchmarks were identified after running all four, so their selection is post hoc; what is pre-specified is the benchmark set, not which members would separate the conditions. IFEval here is our regex-tokenizer substitute for the official harness (Appendix B), applied identically to every condition, so it supports paired comparisons but its absolute values are not directly comparable to published IFEval scores. Second, and more decisively, the single delta does not merely saturate, it *turns over*. Its compliance peaks at $\lambda = 1$ (0.761) and then falls — 0.744 at $\lambda = 1.5$, 0.692 at $\lambda = 2$, 0.516 at $\lambda = 3$ — while

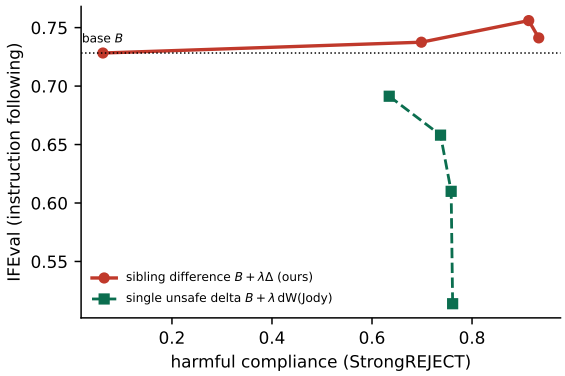

Figure 2: Instruction-following against harmful compliance as each edit is scaled on Llama. Scaling the single unsafe delta trades IFEval away and stalls below 0.77 compliance; the sibling difference reaches 0.91 with IFEval at or above base. MMLU and ARC do not separate the two (Table 2).

capability collapses (-24.2 and -41.8 GSM8K). We swept past the point of usefulness precisely so that the comparison does not rest on a truncated baseline: no scaling of the unsafe adapter reaches the dial's 0.913, because scaling it further destroys the model that would have to comply. The difference reaches it at IFEval +2.8. Figure 2 shows the two frontiers directly: the dial's is not merely lower, it is longer — the sibling difference reaches a regime the single delta cannot reach at any scale. The two sibling-free edits behave differently from each other and neither dominates the dial. A harmful-only LoRA reaches 0.764 compliance but is the most destructive condition we measured (-25.9 IFEval, -8.3 GSM8K). Weight-space ablation of the refusal direction is cheap but weak here: orthogonalizing every residual-writing matrix (attention output, MLP down-projection and the embedding) against the direction, for each of 6 candidate directions taken from layers 8–28 and reported at the best of them (layer 12), reaches 0.390 compliance at +1.5 IFEval. Published abliterations of this model reach higher compliance than we do, so we treat this as a floor on the edit rather than its ceiling; as a check that the implementation is not simply broken, the inference-time version of the same ablation — a hook, independently implemented — peaks at 0.461 on the same prompts and judge, i.e. the same regime. What our number establishes is not that ablation is weak in general but that a refusal direction read from B's activations, applied either way, did not reach the compliance the sibling difference reaches in our hands. Its activation-space counterpart is the strongest baseline here — steering along the refusal direction reaches 0.812 compliance — but at that point it pays -9.9 GSM8K and -4.1 IFEval, while the sibling difference exceeds its compliance (0.913) with both benchmarks at or above base (+3.6 and +2.8). So on the two benchmarks that separate these conditions at all, the dial is above base where every baseline that reaches comparable compliance is below it.

5.3 Uncertainty over pole training, not just over items

The intervals above resample test items within a single trained pole pair, which is the smaller of the two sources of variance: retraining the siblings changes Δ itself. We therefore repeat the whole construction with 3 independently trained pole pairs (same data, same hyperparameters, different seed) and report the spread at the operating point and at full strength. Harmful compliance at $\lambda^\star$ is 0.621, 0.699 and 0.787 (mean 0.702), and at $\lambda = 1$ it is 0.913, 0.914 and 0.949 (mean 0.925). Every pair reaches the high-compliance regime at full strength and none loses instruction-following at the operating point (IFEval is between +0.9 and +3.0 points across pairs, and no benchmark regresses by more than 1.0 point in any pair), so neither the effect nor the capability result is an artifact of one training run. The seed-to-seed range is nonetheless wider than the item-level intervals — at the operating point compliance spans 16.5 points — so we treat differences smaller than that range as unresolved. This is why we do not read the small GSM8K and IFEval *gains* at the operating point as capability improvement: they are within the training-seed range, and the claim we make is only that capability does not regress. Per-pair capability is in Appendix D.

Table 3: The dial Δ computed once on B, merged onto fine-tunes of B (StrongREJECT before and after). Top: jodify with $+\Delta$; bottom: restore refusal with $-\Delta$.

derivative	operation	harmful (before → after)	over-refusal after
ToolACE-8B (tool use)	$+\Delta$	0.014 → 0.688	0.02
MedSSS (medical)	$+\Delta$	0.015 → 0.913	0.03
abliterated	$-\Delta$	0.934 → 0.675	0.00
GeomLlama (drifted)	$-\Delta$	0.435 → 0.209	0.02
same target (ToolACE), alternative edits computed on B			
refusal-direction weight ablation	B's refusal dir.	0.014 → 0.064	–
harmful-only LoRA	+task vector	0.014 → 0.738	–

5.4 Transfer to fine-tunes of B, without retraining

Because Δ is a weight vector for B's architecture, it can be merged onto any fine-tune of B. Table 3 applies the same Δ, computed once on B, to independently released derivatives. On a tool-use and a medical fine-tune it raises harmful compliance from near zero to 0.688 and 0.913. Reversed, $-\Delta$ lowers harmful compliance in an abliterated (uncensored) derivative from 0.934 to 0.675 and in a capability-fine-tune that had lost safety from 0.435 to 0.209. Capability on the transfer targets is kept: jodifying ToolACE leaves its MMLU at 0.602 (from 0.606) and GSM8K at 0.724 (from 0.732), and MedSSS's MMLU is 0.642 (from 0.516). This is a safety edit, not reversion to B: the full dial keeps ToolACE's GSM8K at 0.724 (base 0.732), whereas merging Jody alone without subtracting Fireman reaches similar harmfulness (0.751) but drops its GSM8K to 0.592; the Fireman subtraction preserves the derivative's own capability. MedSSS's MMLU rise to 0.642 (near B's 0.657) might instead be read as reversion toward B; but were the merge simply reverting MedSSS to B, its harm would return to B's 0.062, whereas it rises to 0.913, far above B. We do not have an account of the capability lift that is consistent with the rest of the paper — Δ has the siblings' shared helpful-data component removed, so it should not be supplying B's general capability — and we leave it unexplained rather than rationalised. What the harm number does establish is the narrower point: the safety change is not reversion to B. One target was not mergeable: on a safety-regularised math model the merged weights lose general capability (MMLU 0.277), so we exclude it and note it here.

A pre-specified transfer study. To avoid resting the transfer claim on hand-picked targets, we also ran it on every surveyed derivative whose *base* harmful compliance is below 0.3 — selection on base harm alone, fixed before any transfer was run — merging the same $+\Delta$ onto each and scoring it with the same judge. Of 23 such derivatives, 13 could be evaluated here (10 are unavailable in our offline environment, and we report that rather than dropping them silently), and 11 of those 13 pass the same 0.5 threshold that defines our operating point: a success rate of 85%, with median harmful compliance rising from 0.024 to 0.806. The 2 that resist are both safety-regularised math fine-tunes from the same lab, which stay near base (≤ 0.04) — the clearest signal we have of what makes a derivative resistant. For comparison, transferring the refusal-direction *ablation* from B to two of these targets reaches 0.085 and 0.208. Per-derivative numbers are in Appendix E.

What re-alignment costs. Restoring refusal is not free either: merging $-\Delta$ into the abliterated model lowers its harmful compliance from 0.934 to 0.675 while its MMLU falls 0.605→0.589 and GSM8K 0.675→0.630. So the defensive direction both under-repairs (a reduction, not a return to B's level) and costs a few points of capability, which is the honest summary of the asymmetry we flagged in Section 5.1.

Transfer, and what the comparison does and does not show. Both alternatives from Section 5.2 can be computed on B and merged onto the same derivative. We ran this on one target, so the following is a single-target observation, not an established general ranking. Ablating B's refusal direction from ToolACE's weights leaves harmful compliance at 0.064 — the edit that reached 0.390 on B itself does not carry over — while the sibling difference reaches 0.688 on the same model. The harmful-only task vector does raise compliance to 0.738, but it takes the derivative's capability with it (MMLU 0.572 and GSM8K 0.637, against 0.602 and 0.724 for the dial). Two caveats bound this. The ablation baseline is transferred as a direction read off B, which is the like-for-like comparison

to our transferred Δ but not what a practitioner would do — recomputing the refusal direction on each derivative requires activations from that derivative, and we did not run it, so the contrast shows that *a direction computed on B does not transfer* rather than that ablation is generally weaker on derivatives. And the head-to-head is $n = 1$: the pre-specified study above establishes that *our* edit transfers broadly, but the same sweep for each alternative is the experiment a ranking would need. A plausible reading of why Δ transfers is that, with the shared fine-tuning damage removed, it is close to orthogonal to what the derivative's own fine-tune changed; we offer that as an interpretation, not a demonstration.

5.5 What the difference removes, and what the safe pole is for

Neither sibling is a clean pole. Fireman's own harmful compliance is 0.519 and Jody's is 0.761; both lose about 3 MMLU and 10–20 GSM8K/IFEval points relative to B, and so does an adapter trained on the benign half of the data alone (Table 2), which locates the capability loss in the shared PKU helpful data rather than in the safety targets. Subtracting a sibling trained on that same helpful data removes the shared component.

What the subtracted sibling has to be. Replacing Fireman with an adapter trained on the benign half of the same data gives nearly the same result (0.880 harmful, capability at base level: MMLU -0.1, GSM8K +2.0, IFEval +2.2), so a *safe* sibling is not what makes the cancellation work. What does matter is that the sibling saw the same helpful data. Subtracting an adapter trained on different benign data (Alpaca) reaches comparable compliance (0.874) but no longer cancels the damage: MMLU -4.5, GSM8K -4.3 and IFEval -17.2 points (Table 2). A difference of two siblings that share *everything*, including the safety targets — two Jody seeds, which differ only in training randomness — leaves behaviour at base level (0.017), as it must if the difference is to carry the target behaviour rather than any adapter-to-adapter variation.

So the operative requirement is a sibling matched on the non-target data, and the safe pole is the instantiation that also yields a usable re-alignment direction ($-\Delta$, Section 5.4), which a benign sibling does not provide. Stated as a recipe: to edit a behaviour without paying a fine-tune's capability cost, train two siblings on the same data differing only in that behaviour's targets, and subtract. This is a sharper claim than the one we set out to test, and the two controls above are what separate it from the weaker readings — that any sibling will do, or that any difference of adapters would.

Direction, not magnitude. A random weight perturbation matched to Δ's per-module Frobenius norm leaves harmful compliance at 0.053, and shuffling Δ's entries within each module (same values, destroyed structure) gives 0.061 — both indistinguishable from base B (0.062), against the dial's 0.913. Both controls also leave capability at base level (MMLU -0.2, GSM8K +0.4, IFEval -0.7; MMLU -0.0, GSM8K +0.2, IFEval +1.1), so they are inert rather than damaging: at this norm a generic perturbation of B changes neither behaviour nor capability, and the dial's effect is a property of where Δ points. Nor is the capability preservation a smaller-perturbation effect: $\|\Delta\| = 18.8$ is within 3% of $\|\mathrm{dW(Jody)}\| = 19.3$, so at equal weight-norm the dial keeps capability where the Jody adapter loses it. At the operating point the perturbation is half that norm; the matched-compliance comparison in Section 5.2 is the test that controls for scale properly.

The two siblings' weight changes are partially aligned (Frobenius cosine 0.53 on Llama, 0.43 on Mistral), consistent with a shared component the difference removes. This geometry does not by itself predict where the effect will be strong: on Qwen3 the cosine is even higher (0.57) and Δ retains a comparable share of the Jody norm, yet the behavioural effect there is small (Section 5.7). A quantity that predicts how much of the safety effect survives the subtraction is open; we do not claim one.

5.6 How far have public fine-tunes drifted?

Building the siblings required a rule-frozen set of 25 public fine-tunes of B (Figure 4). Several have lost much of B's refusal behaviour without any adversarial intent: a medical/geometry model reaches 0.435 and a Basque fine-tune 0.122, against B's 0.062. Our own LoRA fine-tunes on harmless data reach 0.130 (Alpaca) and 0.376 (Dolly); adding 1% harmful examples gives 0.744.

Table 4: The dial $B + \lambda^\star \Delta$ at each model's operating point on three base models: harmful compliance before and after, and capability change vs. each base. The effect is large on Llama and Mistral and much smaller on Qwen3, where compliance still rises above base but reaches only 0.145 against the Jody adapter's 0.759. At the operating point the only significant capability regression on any of the three is MMLU, by at most 1.1 points.

base model	$\lambda^\star$	harmful (base)	harmful (dial)	ΔMMLU	ΔARC	ΔGSM8K
Llama-3.1-8B-Instruct	0.50	0.062	0.699	-0.3	+0.3	+2.6
Mistral-7B-Instruct-v0.3	0.50	0.474	0.901	-0.7	-1.1	-1.1
Qwen3-8B	1.00	0.019	0.145	-1.1	-1.1	+0.0

5.7 Generality across base models

The construction replicates on a second family: training a fresh sibling pair on Mistral-7B-Instruct-v0.3 and forming the same difference raises harmful compliance from 0.474 to 0.901 at $\lambda^\star = 0.5$ (Table 4), with the operating point selected by the same dev-harm rule. At that point the only significant capability regression on either family is MMLU (Llama -0.3 [-0.7, +0.1], not significant; Mistral -0.7 [-1.2, -0.2]); ARC, GSM8K and IFEval are within noise or improve (95% paired bootstrap). Dialling Mistral to full strength $\lambda = 1$ costs 4–11 points, so the capability result is a property of the operating point rather than of the edit at any strength.

On Qwen3-8B the construction reproduces only weakly. The direction is right and the response to λ is monotone — compliance goes 0.019 at base, 0.025 at $\lambda = 0.5$ and 0.145 at $\lambda = 1$, a 7.7-fold rise over base — and capability is preserved as on the other two models (MMLU -1.1 [-1.5, -0.7]). But it never reaches the 0.5 operating threshold, and the Jody adapter alone on the same base reaches 0.759, so most of the safety effect is removed along with the shared damage rather than surviving the subtraction. We report this as an attenuated replication rather than a failure: the mechanism points the same way on all three models and differs in how much of the effect survives, and the weight geometry we can measure does not tell us in advance which regime a base model is in (Section 5.5).

6 Limitations

Our evidence is strongest on Llama-3.1-8B, with full replication on Mistral-7B and an attenuated one on Qwen3 (Section 5.7), and the transfer results are within a base model's own family, since a weight delta is tied to one architecture. The jodify direction is tested on a pre-specified set of derivatives (13 of 23 evaluated; the rest were unavailable in our offline environment), while the restoration direction and the head-to-head against alternative edits rest on fewer targets. Retraining the siblings shifts operating-point compliance by 16.5 points, which bounds how finely the capability differences should be read. Our implementations of the two published baselines are swept over layers but not exhaustively tuned. Harmfulness is judged by two language models that agree closely (Cohen's $\kappa = 0.80$), though a human audit would strengthen it. Finally, the two directions are not symmetric: the make-unsafe edit is the stronger one, and re-alignment gives a partial reduction rather than a repair. Accordingly we release the recipe, the evaluation code and all scored generations, but no jodified weights.

7 Conclusion

Two siblings trained on shared helpful data and opposite safety targets give a single weight vector that edits refusal. Against the simpler edits it is meant to improve on — the same unsafe delta scaled down, a harmful-only fine-tune, and weight-space refusal ablation — it is the only one that reaches high compliance at all, and it keeps instruction-following where they lose 7–26 points. It is computed once on the base model and transfers to other parties' fine-tunes without retraining. What makes the cancellation work is that the subtracted sibling is data-matched, not that it is safe. How much of the safety effect survives that subtraction varies by base model — large on two of three, small on the third — and predicting which is open.

REPRODUCIBILITY STATEMENT

All models, datasets and judges are public: the base models (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Qwen3-8B), PKU-SafeRLHF for sibling training, JailbreakBench, AdvBench, XSTest, MMLU, ARC-Challenge, GSM8K and IFEval for evaluation, and the StrongREJECT rubric (Qwen3-32B) and HarmBench classifier as judges. Section 3 and Section 4 give the sibling construction (LoRA rank 64 on all linear projections, 2 epochs, learning rate 10^{-4}, identical data except the safety targets), the exact merge, and the dev-only rule that fixes $\lambda^{\star}$. Generation is greedy (256 new tokens), so outputs are deterministic given the weights. The supplementary code builds every data split with fixed seeds, trains the siblings, merges, generates, judges and scores; every number in the paper is rendered from a single results file produced by one script, and the frozen protocol with its dated deviations is in Appendix B. The 25 public fine-tunes were selected by a keyword rule frozen before scoring, and their identifiers are listed in the supplement. Experiments ran on NVIDIA B200 GPUs; each sibling trains in about 11 minutes on one GPU.

ETHICS STATEMENT

This work studies how to remove and restore refusal behaviour, which is dual-use: the dial is a more effective attack than defence (Section 5.4). We judged that publishing is net-beneficial because refusal removal from open weights is already easy with existing methods (e.g. abliterated models are public, Section 5.6), while the defensive side of what we provide — a survey method for measuring how far released fine-tunes have drifted (Section 5.6) and an edit that partially re-aligns an unsafe derivative without retraining (Section 5.4) — is usable by maintainers and auditors of open-weight ecosystems. We do not claim a weight-space auditing score: the drift measurement we report is behavioural. To limit misuse we release the code, prompts, judge scores and generations needed to verify our claims, but not any jodified weights or the Jody adapters. Harmful prompts come from existing public benchmarks; no new harmful content was elicited from people, and the study involved no human subjects or personal data.

AI USE STATEMENT

The research questions, ideas, and experimental design are the authors' own. AI tools provided code assistance and debugging, including implementing and running training, generation, evaluation and statistical computations, and writing assistance with drafting and editing. The evaluation code uses the LLM judges described in Section 4. The authors directed and verified this work and take full responsibility for every claim and for any errors introduced by these tools; no LLM is an author.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. Language models are Homer Simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of ACL*, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. Safety arithmetic: A framework for test-time safety alignment of language models by steering parameters and activations. In *Proceedings of EMNLP*, 2024.

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. In *International Conference on Learning Representations*, 2022.

Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung-yi Lee. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In *Proceedings of ACL*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Jiaming Ji, Donghai Hong, Borong Zhang, et al. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. *arXiv preprint arXiv:2406.15513*, 2024.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. LoRA fine-tuning efficiently undoes safety training in Llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

Mantas Mazeika, Long Phan, Xuwang Yin, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL*, 2024.

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. In *Proceedings of NAACL*, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, et al. A StrongREJECT for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. In *arXiv preprint arXiv:2308.10248*, 2023.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operation. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A Activation-space baselines and the frozen protocol

The activation-space methods in Figure 1, and the protocol deviations logged in Appendix B, come from a companion study of the same siblings that we summarise here so that both are self-contained. That study asked whether a model's *internal states* reveal how far it has moved toward the unsafe sibling, and it supplies the inference-time baselines we compare the weight edit against.

Directions. On a set of held-out harmful prompts (the *axis prompts*: the 315 PKU harmful prompts never used in training), we read last-prompt-token residuals at every layer and build one direction per method: *sibling direction* (mean Jody residual minus mean Fireman residual), *single pole* (Jody minus B), *prompted poles* (the same contrast induced by system prompts rather than weights), *response contrast* (states after prefilling 32 tokens of the safe vs. unsafe reference response), and the *refusal direction* (difference of means over harmful vs. harmless prompts, Arditi et al., 2024). Every direction is oriented so larger means more unsafe. A *null axis* built from two Fireman seeds, which differ only in training randomness, measures how much of any direction is seed fingerprint rather than safety.

Layer selection, and why the pre-registered rule was replaced. The registered rule picked each method's layer by maximum d' between its two contrast sets. On the reference data this rule selects adapter fingerprints: at layer 4 the null axis scores a *higher* d' than the sibling contrast, and at the d'-selected layer the sibling direction is not reproducible across pole seeds. We therefore replaced it, before scoring any model, with an outcome-free reproducibility rule (agreement between directions built from two independent seeds, or from two halves of the axis prompts), and we report results under both rules. This is the change logged in Appendix B, and it is why that log refers to d', axis prompts, a null axis and a claim labelled PRIMARY_1: those terms belong to this companion analysis. The steering strengths used for the baselines in Figure 1 are each method's published or dev-selected coefficient, swept over layers but not re-tuned per layer.

B PROTOCOL DEVIATIONS

- 2026-09-24, before scoring: public derivatives whose names differ only by a seed suffix (‘_s43‘/‘_s44‘) are the same recipe and are collapsed to one recipe point for PRIMARY_1 (seed-averaged), as the recipe-level rule requires. Affected: Neelectric SafeGrad_mathv00.07 (s43, s44) and SFT_mathsp_ewc_v00.15 (s43, s44).
- 2026-09-24, before scoring: model-card labels are produced by the frozen keyword rule in scripts/freeze_public_derivatives.py and are not hand-edited, including borderline cases.
- 2026-09-24, before scoring: ‘axis_prompts‘ are the 315 held-out PKU harmful prompts (pole_dev_harmful; never trained on), not 500 — only 315 remain after deduplication. The same prompts serve the behavioural stopping rule (judged by Qwen3Guard). No derivative outcome touches them.
- 2026-09-24, before scoring: response_contrast ”matched position” is implemented as the last token after prefilling the first 32 tokens of the safe vs unsafe PKU response on axis prompts, i.e. the state from which the model would continue a response, matching the last-prompt-token read position of every other method.
- 2026-09-24, before any derivative is scored: the pole stopping rule was not met by either pole on pole_dev. Qwen3Guard unsafe rate on the 315 held-out PKU harmful prompts, Llama SFT seed 0, steps 100/200/300/400/492: Jody 0.727/0.756/0.762/0.800/0.800 (rule ¿= 0.90); Fireman 0.390/0.400/0.375/0.349/0.349 (rule ¡= 0.05); Fireman over-refusal 0.000 at every step. Cause: PKU-SafeRLHF ”safe” responses frequently engage with borderline requests rather than refuse, and Qwen3Guard counts a share of them as unsafe, so the Fireman threshold was not attainable with this data; the rule was set without checking that. Handling, fixed before scoring: the rule is not re-tuned; both poles are used at the final step (492), which is an equal training budget, and their measured rates are reported. On the evaluation distribution (judge_plain, n=128 smoke subset) Jody's rates are Qwen3Guard 0.906, HarmBench 0.805, StrongREJECT 0.728, against B's 0.023 / 0.037 / 0.051.
- 2026-09-24, before those models are scored: two frozen public derivatives (MohamedAhmedAE Medical-merged, gaparecido financial-reasoner) have tokenizer configs naming a class this transformers version lacks. They are loaded with B's tokenizer (same vocabulary) and their own chat template from tokenizer_config.json.
- 2026-09-24, BEFORE any derivative is scored (only B, Fireman and Jody references had been captured): the pre-registered layer rule (max d' between a method's two contrast sets) is replaced as the PRIMARY rule, because the reference data show it selects adapter fingerprints rather than the safety difference. Evidence (Llama, axis prompts): d' of the

null axis (Fireman seed 0 vs seed 1) exceeds d' of the poles at layer 4 (31.8 vs 27.9), and at the d'-selected layer 5 the Jody-minus-Fireman direction is not reproducible across seeds (cos(J0-F0, J1-F1) = -0.24 at layer 4, -0.12 at layer 6), whereas it is at layers 16-30 (0.75-0.85). New primary rule, still outcome-free: each method's layer maximises the reproducibility of its direction — for pole-based methods (poles, single pole) the cosine between the direction built from seed 0 and from seed 1; for prompt-contrast methods (prompted poles, response contrast, refusal direction), the cosine between the directions built from two random halves of the axis prompts. The null axis uses the poles' layer. Results under the registered d' rule are reported as well, labelled as such.

- Refinement, same time, still before any derivative is scored: the d' failure is specific to contrasts between different adapters (seed fingerprint). Contrasts computed within B's own weights (prompted poles, response contrast, refusal direction) keep the registered d' rule; split-half reproducibility is ˜1.0 at every layer for them and cannot choose a layer. Resulting layers: poles 29, single pole 24, null axis 29 (= poles), prompted 13, response contrast 28, refusal direction 13.
- 2026-09-24, before scoring: gaparecido financial-reasoner repo ships full merged weights plus a stray adapter_config.json; loaded from a local mirror of the snapshot without the adapter files, i.e. the published full weights.
- 2026-09-24, before scoring: identity control furiosa-ai/Llama-3.1-8B-Instruct does not load (repo config error) and is dropped; four identity controls remain. Llama-3.1-Swallow-8B-v0.5 has no chat template; its native-template generation and capture use the plain 'User:/Assistant:' fallback.
- 2026-09-24, before the claim-1 analysis is final: score orientation convention made explicit — every method is oriented so that larger = more unsafe according to its own construction: refusal direction and refusal logit are negated (a larger projection / log-odds means more refusal). This follows from each method's definition, not from outcome data.
- 2026-09-24: IFEval's NLTK punkt data cannot be fetched offline (NLTK refuses proxied downloads). Its sentence splitter and word tokenizer are replaced by regex equivalents inside cap_eval.py, applied identically to every model and steering condition, so paired capability differences are unaffected; absolute IFEval numbers may differ slightly from the official implementation.
- 2026-09-24, found while writing: the steering dev split was the first 64 judge_plain and 64 over-refusal prompts, which are also in the full test set. Test numbers are reported on the remaining judge_plain (n=212) and over-refusal prompts only, i.e. excluding every dev prompt. No operating point changes.

C THE DIAL SWEEP

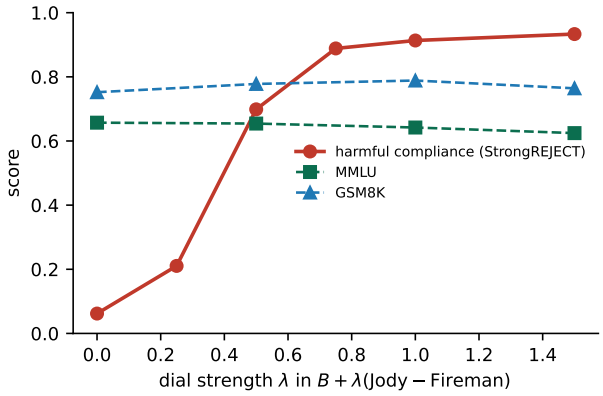

Figure 3: Harmful compliance rises with λ while MMLU and GSM8K stay flat (Llama, seed 0).

D PER-SEED RESULTS AT THE OPERATING POINT

Table 5: The same construction repeated with independently trained pole pairs, at the operating point $\lambda^\star = 0.5$ (harmful compliance, $n = 212$; capability change vs. B in points). Seed-to-seed spread bounds how finely these numbers can be read.

pole pair	harmful	ΔMMLU	ΔARC	ΔGSM8K	ΔIFEval
pole pair 0	0.699	-0.3	+0.3	+2.6	+0.9
pole pair 1	0.787	-0.4	-1.0	+0.4	+2.4
pole pair 2	0.621	+0.5	+0.7	-0.3	+3.0

E PRE-SPECIFIED TRANSFER STUDY: PER-DERIVATIVE OUTCOMES

Every surveyed derivative with base harmful compliance below 0.3, with the same $+\Delta$ merged onto it (selection on base harm alone, fixed before running). "not evaluated" means the weights were not retrievable in our offline environment; no outcome is omitted.

derivative	harmful (base)	harmful ($+\Delta$)
DeepMount00/Llama-3.1-8b-ITA	0.024	0.943
arcee-ai/Llama-3.1-SuperNova-Lite	0.024	0.941
HiTZ/Latxa-Llama-3.1-8B-Instruct	0.122	0.941
pixas/MedSSS_Policy	0.015	0.926
NbAiLab/nb-notram-llama-3.1-8b-instruct	0.047	0.918
Kukedlc/NeuralLLaMa-3-8b-ORPO-v0.3	0.085	0.810
Team-ACE/ToolACE-8B	0.014	0.806
MohamedAhmedAE/Llama-3.1-8B-Instruct-Medical-F	0.025	0.786
CNCL-Penn-State/CrPO-sft-llama-3.1-8b-instruct	0.097	0.756
watt-ai/watt-tool-8B	0.023	0.746
SentientAGI/Dobby-Mini-Unhinged-Llama-3.1-8B	0.111	0.655
Neelectric/Llama-3.1-8B-Instruct_SafeGrad_math	0.000	0.040
Neelectric/Llama-3.1-8B-Instruct_SafeGrad_math	0.000	0.035
gaparecido/llama-3.1-8b-financial-reasoner-v1	0.054	not evaluated
Neelectric/Llama-3.1-8B-Instruct_SFT_mathsp_ew	0.043	not evaluated
ApolloRaines/Llama-3.1-8B-Instruct-Security-An	0.009	not evaluated
Neelectric/Llama-3.1-8B-Instruct_SFT_mathsp_ew	0.032	not evaluated
ApolloRaines/Zark-Muckerberg-Llama-3.1-8B	0.082	not evaluated
Neelectric/Llama-3.1-8B-Instruct_SFT_mathsp_ew	0.073	not evaluated
tokyotech-llm/Llama-3.1-Swallow-8B-v0.5	0.156	not evaluated
Vikhrmodels/Vikhr-Llama3.1-8B-Instruct-R-21-09	0.016	not evaluated
barc0/Llama-3.1-ARC-Potpourri-Induction-8B	0.030	not evaluated
ApolloRaines/Llama-3.1-8B-Instruct_Deep-Analys	0.059	not evaluated

F PUBLIC DERIVATIVES

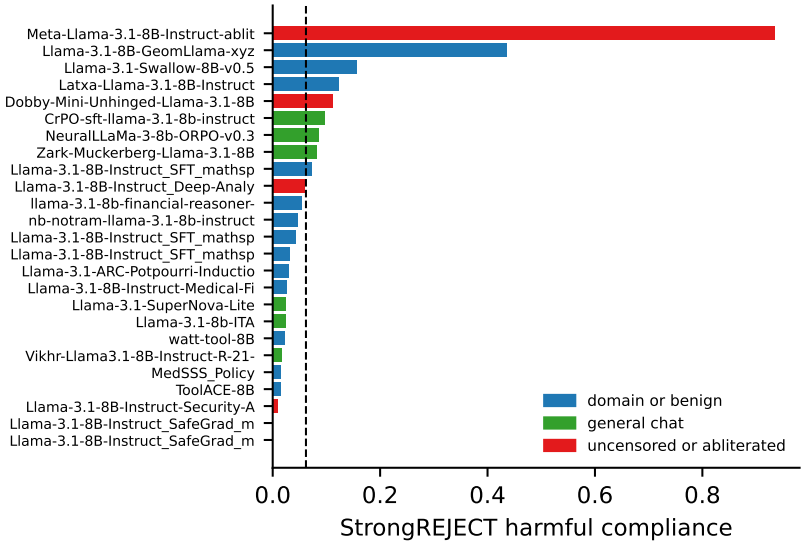

Figure 4: Harmful compliance of the 25 rule-frozen public fine-tunes of B; dashed line: B.